ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 26, No. 11, pp. 493-498, 2020. 11
https://doi.org/10.5626/KTCP.2020.26.11.493

# 음악 유사도 비교를 위한 Siamese 네트워크 기반 그래프 임베딩의 개선

## (Improvement of Graph Embedding Based on Siamese Network for Comparison of Music Similarity)

송 창 헌 † 이 용 현 † 김 형 주 ††
(Changheon Song) (Yonghyun Lee) (Hyungjoo Kim)

**요 약** 음악 시장의 성장에 따라 사용자는 일부 음악에 국한되어 노출되고 선택하게 된다. 많은 서비스는 메타데이터로 라이브러리를 구성하여 검색 및 추천 문제에 접근하고 있다. 이때, 새로 나오거나 인지도가 없는 음악의 경우 결과에서 제외될 수 있다. 일반적으로 사용되는 오디오 피처는 해상도에 따른 차원의 변화 폭이 크기 때문에 CNN의 입력으로 사용하기에 어려움이 있다. 본 논문에서는 음악 그래프 피처를 추출하고 임베딩하여 유사도를 비교할 수 있는 모델을 제안한다. 모델은 피처 추출과 Siamese 네트워크로 구성된다. 피처 추출에서는 각 음악 신호를 오디오 피처로 변환하고, 각 음악의 그래프 피처를 구성한다. 이후, Siamese 네트워크에서 각 그래프 피처를 GCN과 어텐션 기법을 활용하여 잠재 공간으로 임베딩하고, NTN을 통해 서로 다른 두 벡터의 유사도를 도출한다. 마지막으로 실험을 통해 음악 신호의 유사도 비교를 위한 오디오 피처의 그래프 피처 추출이 효과적인 방식임을 입증하였다.

**키워드:** 오디오 컨텐츠, 그래프 임베딩, 그래프 콘볼루셔널 네트워크, Siamese 네트워크

***Abstract*** As the music market grows, people are exposed to and provided with selective music. Many services use metadata for building music libraries. In this situation, songs from independent labels and new artists that do not have previous information are still excluded from the result of searches and recommendations. In this paper, we focus on making the music scoring model for calculating the similarity score of two music signals. The model comprises the Siamese network and the scoring layer. The Siamese network embeds audios to small latent vectors and passes them to the scoring layer. The audio feature is difficult to use as an input to the CNN because of the dimensionality problem. Our approach is compared to previous works because it retains the sequence information of the peak frequencies in the spectrogram by transforming it into a graph. The effectiveness of the graphical approach is shown as the result of the experiment.

**Keywords:** audio content, graph embedding, graph convolutional network, Siamese network

· 이 논문은 2020 한국연구재단의 재원으로 서울대학교 컴퓨터공학부 BK21 FOUR 지능형컴퓨팅사업단의 지원을 받아 수행된 연구임(4199990214639)

† 비 회 원 : 서울대학교 컴퓨터공학부 학생(Seoul Nat'l Univ.)
chsong@idb.snu.ac.kr
(Corresponding author임)
leeyh@idb.snu.ac.kr
†† 종신회원 : 서울대학교 컴퓨터공학부 교수
hjk@snu.ac.kr

논문접수 : 2020년 7월 7일
(Received 7 July 2020)
논문수정 : 2020년 9월 15일
(Revised 15 September 2020)
심사완료 : 2020년 9월 18일
(Accepted 18 September 2020)


정보과학회 컴퓨팅의 실제 논문지 제26권 제11호(2020. 11)

## 1. 서 론

음악 시장이 성장함에 따라, 많은 수의 음악과 앨범이 Apple music, Youtube Music, 그리고 Spotify와 같은 스트리밍 서비스나 온라인 음반 판매 서비스를 통해 발매되고, 이를 소비하는 사람도 늘어나고 있다. 많은 수의 음악 서비스는 주로 음악의 메타데이터를 미리 구성하여 저장 및 관리하며, 이 데이터를 활용하여 검색 또는 추천을 제공한다. 하지만 메타데이터를 이용하여 음악 라이브러리를 구성하는 경우에는 음악의 컨텐츠를 기반으로 하는 어플리케이션의 경우 어려움을 가진다.

예를 들어, 서비스에서 특정 분위기를 지닌 음악과 비슷한 음악을 하나의 플레이리스트로 구성하여 제공하는 경우, 모든 음악의 분위기를 기억하거나 추가할 음악과 미리 구성된 플레이리스트의 모든 음악을 반복해서 들어야 한다. 또 새로 발매되거나 인지도가 없는 가수의 음악인 경우, 미리 정의된 정보가 없어 추천과 검색결과에서 제외될 수 있다.

또한, 음악의 양이 많아짐에 따라 사용자는 선택적으로 음악을 소비하게 되고, 서비스는 각 사용자에게 맞게 추천을 해야 할 필요성이 있다. CF(Collaborative Filtering) 방법은 사용자에게 맞는 아이템을 추천함에 있어 높은 정확도를 보이고, 많은 서비스가 채택해서 사용하고 있다[1]. 이는 사용자들의 유사성과 노래들의 유사성을 기반으로 추천해줄 음악에 대한 점수를 계산하는 방식이다. 이를 음악 시장에 접목하여 각 음악 사이의 유사성을 계산할 때, 전문가에 의해 미리 정의된 메타데이터들을 이용하려면 메타데이터 구성에 있어 도메인 지식을 필요로 한다. 이 과정에서 음악의 신호, 즉 오디오 컨텐츠가 지니는 일부 특성을 잃게 된다. 따라서 본 논문에서는 추가적인 메타데이터 없이 음악 유사도를 측정하는 모델을 제안한다.

### 1.1 관련 연구

서로 다른 음악의 유사도 측정에 있어 오디오 피처를 활용하는 연구가 기존에도 많이 진행되어왔다. Slaney et al.[2]와 Schluter et al.[3]은 오디오 신호의 유사성을 파악하기 위해 적절한 Similarity Metric에 관해 연구 하였다. 두 연구에서는 적절한 Similarity Metric을 찾기 위해서 도메인 지식을 활용하였다. Logan et al.[4]은 오디오 신호에서 얻은 MFCCs(Mel-Frequency Cepstral Coefficients)의 스펙트럼 거리(Spectral Distance)로 두 음악 사이의 유사도를 계산하였다.

최근 연구에서는 오디오 컨텐츠가 결과에 영향을 미칠 수 있도록 오디오 피처를 입력으로 신경망을 구성하는 연구도 진행되었다. Wang et al.[5]은 스펙트로그램을 Deep Belief Network의 입력으로 사용하여 음악을 추천할 수 있도록 했다. MFCCs를 오디오 피처로 활용하고 CNN(Convolutional Neural Network)와 WMF(Weighted Matrix Factorization)를 사용하여 오디오 컨텐츠의 잠재적인 벡터를 예측하는 연구도 있었다[6]. Zhang et al.[7]은 따라 부른 곡과 원곡의 유사성을 비교하기 위해 CQT를 통해 얻은 스펙트로그램을 활용하여 CNN을 통해 둘 사이의 유사성을 판단하는 모델을 제시하였다.

신경망을 활용하여 오디오 피처를 입력으로 사용하는 방식에 대해 Korzeniowski et al.[8]은 오디오 프레임 수준에서 복잡한 시간 모델 학습하는 것은 좋은 성능을 보이지 못한다고 언급하였다. CQT변환을 통해 스펙트로그램을 형성할 때, 샘플링 레이트(Sampling Rate)에 따라 오디오 피처는 시간 축에 대해 고차원의 결과를 보인다. 이를 피하기 위해 샘플링 레이트를 낮추어 진행하는 경우 음악의 해상도(Resolution)가 떨어지는 문제가 발생한다. 이러한 문제를 해결하기 위해 본 논문에서는 샘플링된 오디오 프레임의 정점을 이루는 주파수 대역을 노드로 하고 이어진 프레임의 정점들과 간선을 가지는 고정된 크기의 그래프를 형성하고 임베딩하여 유사도를 구하는 모델을 제안한다.

최근 그래프 구조를 임베딩하는 방법에 관한 연구가 활발히 진행되고 있다. Kipf et al.[9]은 스펙트럴 그래프 콘볼루션(Spectral Graph Convolution)을 통해 그래프 구조를 학습하는 모델을 제안하였다. simGNN [10]에서는 GCN(Graph Convolution Network)과 어텐션(Attention) 기법을 융합하여 그래프 임베딩을 도출하고, 노드 쌍 사이의 비교(Pairwise Comparison) 정보를 추가하여 두 그래프 임베딩 사이의 유사도를 구하였다. 본 논문에서는 두 음악 그래프를 임베딩하고 유사도를 비교 하는 모델 구축을 위해 다양한 그래프 임베딩 방법 중 SimGNN에서 제안된 Siamese 네트워크 기반 그래프 임베딩 방식을 활용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 제안하는 모델을 오디오 그래프 추출 모델과 임베딩 및 점수 계산 모델로 나누어, 전체 모델 구성에 관해 서술한다. 3장 에서 제안하는 모델을 시험해보기 위해 사용한 데이터에 관해 기술하고 매개 변수를 변경하여 실험한 결과를 도식화하여 분석하고 서술한다. 마지막 4장에서는 실험 결과를 바탕으로 제안하는 모델에 대한 결론을 도출한다.

## 2. 제안 모델

본 논문에서 제안하는 모델은 오디오 그래프 피처 추출 모델과 그래프 피처를 잠재 공간에 임베딩하고 유사도를 측정하는 점수 계산 모델로 구성된다. 이에 대한 전반적인 구조는 그림 1을 따른다.

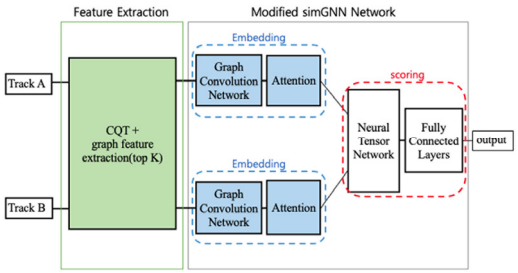


그림 1 제안 모델 구조

Fig. 1 Overview structure of the proposed model

### 2.1 오디오 그래프 추출 모델

오디오 피처를 그대로 활용하는 기존의 방식과 다르게, 피처 추출 모델에서는 음악 신호로부터 그래프 피처를 형성한다. 이에 대한 전반적인 과정은 그림 2를 따른다.

오디오 피처를 구성하기 위해서 사용하는 푸리에 변환은 신호를 분해하여 일정 시간 동안 각 주파수의 영향력을 파악할 수 있도록 스펙트로그램을 결과로 도출한다. 오디오 신호를 스펙트로그램으로 변환하는 경우,

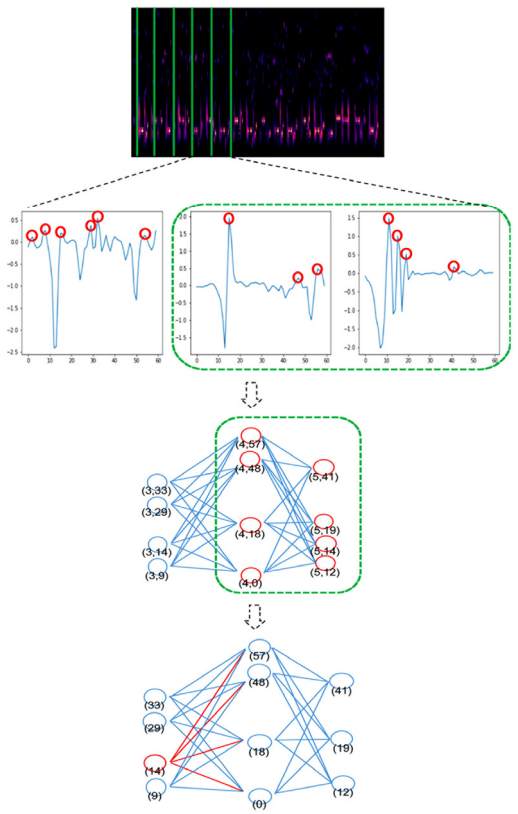


그림 2 오디오 그래프 피처 추출 과정

Fig. 2 Process of the audio graph feature extraction

변환을 위한 샘플링 레이트에 따라 스펙트로그램을 구성하는 오디오 해상도가 달라진다. 본 논문에서는 모든 음악을 동일한 샘플링 레이트를 적용하여 같은 해상도의 스펙트로그램이 나올 수 있도록 구성했다. 해상도가 증가할수록 음악 신호의 특성을 자세하게 내포할 수 있으나, 오디오 피처의 차원이 증가하여 신경망 활용에 어려움이 있다. 이를 해결하기 위해, 하나의 음악에 대해 분해된 주파수 또는 주파수 대역의 정점(Peak)을 노드로 구성하는 방법을 제안한다. 이때 한 프레임 속의 주파수가 많을수록 그래프의 크기가 증가하게 되는데, 본 논문에서는 계산의 효율성을 위해 푸리에 변환의 변형인 CQT(Constant-Q Transform)를 사용하여 일정 범위의 주파수를 하나로 묶어주도록 구성하여 그래프가 적은 수의 노드로 구성될 수 있도록 한다. 이후, 스펙트로그램의 이웃한 프레임의 주파수 대역들을 간선으로 연결하여 그래프로 구성하고 이를 신경망의 입력으로 활용하는 방법을 제안한다.

그래프를 구성할 때, 오디오 피처의 시간 축에 대한 선형 샘플링(Linear Sampling)을 통해 일부 프레임을 추출하여 계산의 효율성을 높였다. 또한 샘플링된 프레임에서 영향도가 정점을 이루는 주파수 대역 중 영향도 값이 큰 순서로 K개를 선택하여 노드를 구성함으로써 그래프의 크기를 줄였다. 이때 샘플링된 프레임의 상황에 따라, 정점의 개수가 K를 채울 수 없는 경우에는 제로 패딩(Zero-Padding)을 통해 차원의 크기를 맞춰주었다. 하나의 음악 신호는 피처 추출 모델을 거쳐 연속된 샘플 프레임의 노드가 서로 연결된 일정한 차원의 그래프 형태를 가진다.

### 2.2 임베딩 및 점수 계산 모델

임베딩 및 점수 계산 모델은 오디오 그래프 추출 모델을 통해 형성된 서로 다른 두 그래프를 입력으로 받아 잠재 벡터로 임베딩하고, 두 벡터의 유사도 점수를 계산한다. 임베딩 및 점수 계산 모델의 전반적인 구조는 그림 1의 Modified SimGNN Network 부분을 따른다.

본 논문에서는, simGNN[10]을 기반으로 두 그래프를 입력으로 받아 두 그래프의 유사도를 예측할 수 있는 모델을 구성한다. 이때 사용된 모델은 다음과 같다. 하나의 그래프가 통과하는 임베딩 타워는 3번의 GCN으로 구성되며 노드 수준의 임베딩을 수행한다. 임베딩 모델은 Siamese 네트워크의 구조를 가지며, 서로 다른 입력 그래프에 대해 같은 임베딩 타워를 활용한다. 이후, 노드 수준의 잠재 벡터로부터 어텐션 기법을 이용하여 그래프 수준의 임베딩을 계산한다.

계산된 두 그래프 수준의 임베딩에 대해 NTN(Neural Tensor Network)을 활용하여 임베딩된 두 그래프 사이의 관계를 나타낼 수 있도록 한다. 두 임베딩된 그래

프가 NTN을 통과하여 상호관계를 표현하는 점수들로 이루어진 벡터를 구성하게 되면, 이를 완전 연결 레이어를 통과시켜 일차원 점수로 만들고 시그모이드(Sigmoid) 활성 함수를 적용함으로써 결과가 유사도 값인 0과 1 사이의 값을 결과로 도출하도록 한다. 이를 통해 얻은 유사도와 실제 두 음악 사이의 유사도의 MSE(Mean Squared Error)를 손실 함수로 사용하여 모델을 학습한다. 모델 학습을 위한 알고리즘은 Algorithm 1에 나타내었다.

**Algorithm 1** Pseudo Code of Embedding and Scoring Model

**for each batch do**
  **for** iter = 1, …, max_iter **do**
    graph features *f1,f2*, edge indices *e1,e2* ← *X*
    **for** each layer *l* to 3**do**
      *f1* = GCNconv(*f1, e1*)
      *f2* = GCNconv(*f2, e2*)
    **end**
    *f1′, f2′* = compute weighted sum of node-level embeddings *f1,f2*
    $S_k$ = compute *K* dimensional interaction scores between graph-level embeddings *f1′,f2′* using *Neural Tensor Network*
    *s* = reduce dimension to 1 of $S_k$ using *Fully Connected Neural Network*
  **end**
  update weight of model
**end**

## 3. 실 험

본 논문에서는 제안하는 모델의 검증을 위해 오디오 그래프 추출과 임베딩 및 점수 계산 모델의 구조를 변경하며 학습을 진행했다. 먼저 임베딩 및 점수 계산 모델 구조 변경 실험을 통해 생성한 음악 그래프 피처에 대해 음악 그래프의 특성을 잘 포함하도록 임베딩할 수 있는 적절한 구조를 찾는 실험을 진행했다. 이후, 본 논문에서 제안하는 방식에 따라 오디오 피처의 그래프 변환이 가지는 장점을 비교해 보기 위해 기존 연구와 비교 실험을 진행했다.

선행하는 연구에서 Zhang et al.[7]은 음악과 따라 부른 음악의 피처를 추출하기 위해 Convolutional Semi Siamese Network인 IMINET을 제안했다. 연구에서 제안하는 모델은 주어진 입력 음악에 대해 CQT를 통해 스 펙트로그램으로 변환하고, CNN과 완전 연결 레이어를 통해 비슷한 정도를 예측하여 비슷하거나 비슷하지 않음으로 판단하여 결론을 내린다.

일반적으로 스펙트로그램은 샘플링 레이트에 따라 차원의 변화가 크기 때문에 신경망 활용에 어려움이 있다. 따라서 본 논문에서는 오디오 피처로부터 그래프를 형성하고, 이를 활용하여 유사도를 비교하는 기법을 제안한다. 본 논문이 제안하는 모델의 성능과 스펙트로그램을 CNN으로 임베딩하는 IMINET의 모델 성능을 비교함으로써 제안하는 기법을 검증했다.

본 논문에서는 모델의 구현을 위해, 실험 환경으로는 Ubuntu 16.04.2 LTS 운영체제, Intel® Xeon ® CPU E5-2620 v4 @ 2.10Ghz CPU, 192GB Samsung DDR4 RAM 5개, GeForece TITAN Xp 12GB 1개를 사용하였 고, Python 3.7과 PyTorch를 활용하였다.

### 3.1 데이터 셋

실험을 위해 사용된 데이터는 MSD(Million Songs Dataset)로, 백만 개의 음악에 대해 미리 계산된 오디오 피처와 메타데이터로 구성돼있는 데이터 셋이다. 음악에 대한 사용자의 선호도 정보와 음악 사이의 유사도 정보 등을 포함하는 Last.fm 등 다른 음악 관련 데이터 셋이 MSD에 연결되어있다. 본 실험에서는 두 음악 사이의 유사성 비교를 위해 Last.fm의 음악 사이의 유사도 데이터를 활용하여 유사도 계산 모델을 학습한다. 데이터 셋은 약 58만 곡의 유사도 정보를 담고 있으나, 활용할 수 있는 자원의 한계로 인하여 21,702개의 곡에 대한 정보를 실험에 활용했고, 이 곡들에 대해 약 65만 쌍의 유사도 점수가 활용되었다. 본 논문에서는 tag나 다른 메타데이터를 제외하고 오디오 신호를 기반으로 실험을 진행하므로 약 30초 길이의 음악 미리 듣기 음원을 활용했다.

### 3.2 실험 모델

음악 그래프 피처의 특성을 내포하도록 임베딩 모델의 구조를 변경하여 진행한 실험은 Siamese 네트워크의 임베딩 타워를 구성하는 GCN의 구조, 점수 예측을 위한 모델의 후반부의 NTN의 차원의 크기, 그리고 마지막 완전 연결 레이어의 차원을 변경하여 실험 결과를 비교하였다. 기존의 그래프 임베딩 연구에서 좋은 결과를 보였던 128-64-32의 구조를 기준으로, 레이어의 차원을 변경하여 64-64-32와 64-32-16의 구조를 가지는 GCN 임베딩 타워를 적용한 모델을 비교하여 실험했다. 또한 NTN의 크기와 완전 연결 레이어를 16, 32, 그리고 64차원으로 변경하여, 음악 그래프의 특성을 보존하며 임베딩하는 모델을 선택하여 총 27개의 서로 다른 구조에 대해 실험하고 비교했다.

이후, 음악 그래프 변환과 활용의 장점을 확인하기 위해 음악 그래프 추출을 위한 매개변수를 변경하여 다양한 형태의 오디오 피처에 대해 실험을 진행했다. 본 논문에서는 그래프 및 오디오 피처의 형태를 변화시킬 수 있는 매개변수로써 CQT 변환 주파수 대역의 개수, 선형 샘플링에 이용한 프레임의 수를 변화시켜 결과를 비교했다. CQT를 이용하여 오디오 신호를 변환할 때, 음

악의 충분한 고저를 포함하기 위해 72개, 60개의 범위를 사용하여 묶는 경우로 나누어 실험을 진행했다. 적용하는 CQT변환의 샘플링 레이트에 따라 시간 축 차원이 달라지기 때문에 22,050Hz의 고정된 샘플링 주파수로 30초 길이의 음원 파일의 오디오 피처를 계산하고, 선형 샘플링을 이용하여 1초 간격과 0.5초 간격을 의미하는 30개 또는 60개의 프레임에 대한 결과를 비교했다. 추가로 그래프의 크기에 영향을 줄 수 있는 프레임별 주파수 대역의 정점을 15개 또는 30개를 선택하는 방식으로 나누어 실험을 진행하고 비교했다.

### 3.3 실험 결과

본 논문에서는 시간 및 자원의 제약으로 총 27가지의 임베딩 및 점수 계산 모델 구조를 실험을 통해 얻은 결과를 비교하여 가장 좋은 성능을 보이는 모델을 선택했다. 60개의 주파수 대역(Bins)과 30개의 선형 샘플링 프레임으로 구성한 오디오 피처를 고정하여 사용하고, 오디오 피처로부터 그래프를 추출할 때는 프레임 별 30개의 정점 값을 고정으로 이용했다.

모델 구조 변경 실험 중, 좋은 성능을 보인 16가지 실험의 결과는 그림 3과 같다. 128-64-32와 64-64-32 의 GCN 임베딩 타워 구조를 가지는 경우, 모델의 NTN 레이어의 차원이 16일 때 가장 낮은 MSE를 보였다. 그러나 일부 모델에서 MSE 값이 큰 폭으로 증가하는 불안정한 결과를 도출했다. 64-32-16의 구조의 모델을 이용한 실험에서는 NTN 레이어의 차원과 관계없이 안정적으로 좋은 성능을 보였다. 실험에 사용한 27가지 모델 구조 중 기존 그래프 임베딩 연구에서 좋은 결과를 가졌던 128 -64-32의 임베딩 타워 구조를 가지며 NTN 레이어의 차원이 16이고 16차원의 완전 연결 레이어를 가지는 모델의 MSE 값이 5.02361로 가장 좋은 성능을 보였다.

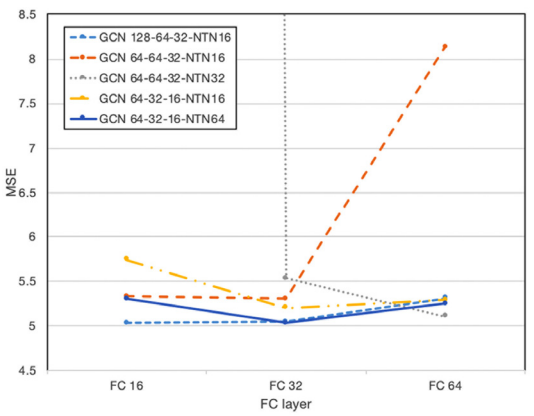


그림 3 임베딩 및 점수 계산 모델 구조 변경 실험 결과
Fig. 3 Experiment result of the changing embedding and scoring model structure

표 1 제안 모델과 IMINET 비교 실험 결과
Table 1 Experiment comparison between the proposed model and IMINET

| FE Parameters | IMINET | Our Model |
|---|---|---|
| 72 bins 30 frames 15 K | 5.44627 | 5.0441 |
| 72 bins 30 frames 30 K | | 245.30802 |
| 72 bins 60 frames 15 K | 7.82277 | 5.36102 |
| 72 bins 60 frames 30 K | | 5.61305 |
| 60 bins 30 frames 15K | 6.60767 | 5.38286 |
| 60 bins 30 frames 30K | | 5.02316 |
| 60 bins 60 frames 15K | 7.08086 | 4.66029 |
| 60 bins 60 frames 30K | | 5.48543 |

이후, 임베딩 및 유사도 점수 계산 모델 구조 변경 실험에서 가장 낮은 MSE 결과를 보인 GCN128-64-32-NTN16-FC16 모델 구조를 활용하여 제안 모델과 기존 연구의 비교 실험을 진행했다. 주파수 대역의 수, 선형 샘플링 프레임 수, 프레임별 정점의 수를 변경하여, 오디오 피처의 그래프 변환 검증 실험을 진행했다. 검증을 위한 비교 모델로는 오디오 피처를 CNN의 입력으로 사용하는 기존 연구로 IMINET 모델을 이용하고, 선행 연구에서 제안한 Conv12-P(2,4)-Conv12-P(2,4)-Conv6-Conv6-FC43 2-FC32 구조를 이용하여 두 오디오 피처의 유사도를 계산했다.

실험 결과는 표 1과 같다. 실험 결과 그래프 형성을 위한 프레임 당 정점의 개수 K가 30인 경우, 전반적으로 오디오 피처를 CNN의 입력으로 사용한 IMINET보다 좋은 결과를 보였다. 그러나 일부 경우에서는 더 많은 더미 노드를 생성하여 MSE가 커지는 현상이 발생했다. 프레임 별 주파수 대역 정점의 수를 줄여 K를 15로 하여 실험을 진행한 경우, 모든 경우에서 안정적인 그래프를 형성하여 더 낮은 오류를 가졌다. 음악의 충분한 고저를 나타낼 수 있도록 변환하기 위한 변수들을 사용하는 경우에 대한 두 모델의 비교 실험을 통해, 프레임별 정점의 수 K를 적절하게 선택하여 더미 노드를 포함하지 않고 그래프를 형성하도록 모델을 구성하면, 제안하는 모델이 기존의 오디오 피처를 CNN의 입력으로 활용하여 유사도를 비교하는 모델보다 음악 신호의 유사한 정보를 더욱 자세히 포함하도록 임베딩할 수 있음을 확인하였다.

## 4. 향후 연구

본 논문에서 오디오 피처로부터 생성된 음악 그래프는 음악의 프레임별 정점의 연속적인 정보를 간선에 담고 있다. 그러나 기존 연구보다 좋은 결과를 보였던 임베딩 및 점수 계산 모델의 임베딩 타워에서 활용한 그래프 임베딩 기법 GCN은 이웃 노드 집합(Neighbor

Aggregation) 기반의 방법으로 간선을 제외하고 집합의 과정이 이루어진다. 향후 연구에서는 임베딩 타워를 구성하는 그래프 임베딩 기법을 간선에 중점을 둔 그래프 집합 기법을 활용하도록 변경하여 실험을 진행하고 비교하고자 한다. 또한 음악 유사도 데이터의 특성에 따라 유사도 값이 존재하지 않거나 유사하지 않은 경우가 상대적으로 많다. 따라서 향후 연구에서 다양한 샘플링 기법을 적용하여 학습 데이터의 불균형을 해결하고 결과를 비교해보자 한다.

## 5. 결 론

메타데이터를 이용한 음악 라이브러리를 구성하면서 새로운 곡 또는 유명하지 않은 곡들이 검색 및 추천 결과에서 제외될 수 있는 문제가 발생한다. 이를 위해 선행하는 연구에서는 오디오 피처를 신경망의 입력으로 이용하여 음악의 컨텐츠를 임베딩하는 방법을 제안했다. 입력으로 사용하는 오디오 피처는 샘플링 레이트에 따라 차원 변화의 폭이 큰 문제점을 해결하기 위해, 본 논문에서는 오디오 피처의 정점들과 연속된 프레임의 관계를 이용하여 그래프로 구성하고, 이를 임베딩하여 유사도를 계산하는 기법을 제안했다.

본 논문에서 제안하는 모델을 사용함으로써, 약 2만 개의 음악에 대한 유사도 측정에서 기존 연구보다 좋은 성능을 보임을 검증하였다. 또한 여러 가지 임베딩 모델 하이퍼 파라미터 변경 실험을 통해 제안하는 모델이 음악 그래프를 속성을 잘 포함할 수 있도록 임베딩할 수 있음을 확인하였다. 검증을 위한 실험에서 제한된 자원의 이용으로 인해 데이터의 일부분을 발췌하여 연구를 진행하였지만, 모든 데이터를 사용하여 임베딩 모델을 학습시킬 경우 더욱 낮은 에러를 보일 것으로 예상되며 추후 연구를 진행할 계획이다. 또한 제안한 모델의 임베딩 타워에서 그래프의 간선에 비중을 두고 임베딩할 수 있도록 다른 방식의 그래프 임베딩을 적용해볼 수 있다.

## References


[1] J. Bennett and S. Lanning, "The Netflix Prize," *Proc. of K DD cup and workshop*, pp. 35, 2007.

[2] M. Slaney, K. Weinberger and W. White, "Learning a Metric for Music Similarity," *Proc. of International Symposium on Music Information Retrieval (ISMIR)*, 2008.

[3] J. Schluter and C. Osendorfer, "Music Similarity Estimation with the Mean-Covariance Restricted Boltzmann Machine," *Proc. of 2011 10th International Conference on Machine Learning and Applications and Workshops*, pp. 118-123, 2011.

[4] B. Logan and A. Salomon, "A Music Similarity Function Based on Signal Analysis," *Proc. of IEEE International Conference on Multimeida (ICME)*, pp. 22-25, 2001.

[5] X. Wang and Y. Wang, "Improving Content-Based and Hybrid Music Recommendation Using Deep Learning," *Proc. of the 22nd ACM international conference on Multimedia*, pp. 627-636, 2014.

[6] A. Van den Oord, S. Dieleman and B. Schrauwen, "Deep Content-Based Music Recommendation," *Proc. of Advances in neural information processing systems*, pp. 2643-2651, 2013.

[7] Y. Zhang and Z. Duan, "Iminet: Convolutional Semi-Siamese Networks for Sound Search by Vocal Imitation," *Proc. of 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 304-308, 2017.

[8] F. Korzeniowski and G. Widmer, "On the Futility of Learning Complex Frame-Level Language Models for Chord Recognition," arXiv preprint arXiv:1702.00178, 2017.

[9] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," arXiv preprint arXiv:1609.02907, 2016.

[10] Y. Bai, H. Ding, S. Bian, T. Chen, Y. Sun and W. Wang, "Simgnn: A Neural Network Approach to Fast Graph Similarity Computation," *Proc. of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 384-392, 2019.



송 창 헌

2019년 국민대학교 컴퓨터공학부 학사 2019년~현재 서울대학교 컴퓨터공학부 석사과정 재학 중. 관심분야는 데이터 마이닝, 컴퓨터 청각, 추천 시스템



이 용 현

2015년 성균관대학교 컴퓨터공학부 학사 2015년~현재 서울대학교 컴퓨터공학부 석박사통합과정 재학 중. 관심분야는 그래프 데이터마이닝, 그래프 컨볼루션 네트워크, 빅데이터



김 형 주

1982년 서울대학교 전산학과 학사. 1985년 Univ. of Texas at Austin 석사. 1988년 Univ. of Texas at Austin 박사. 1988년~1990년 Georgia Institute of Technology 조교수. 1991년~현재 서울대학교 컴퓨터공학부 교수. 관심분야는 데이터베이스, XML, 시맨틱웹, 빅데이터