특집논문 (Special Paper)
방송공학회논문지 제24권 제6호, 2019년 11월 (JBE Vol. 24, No. 6, November 2019)
https://doi.org/10.5909/JBE.2019.24.6.1024
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

분리된 보컬을 활용한 음색기반 음악 특성 탐색 연구

이 승 진[a)‡]

Investigation of Timbre-related Music Feature Learning using Separated Vocal Signals

Seungjin Lee[a)‡]

요 약

음악에 대한 선호도는 다양한 요소들에 의해 결정되며, 추천의 이유를 보여주는 특성을 발굴하는 것은 음악 추천에 있어 중요하다. 본 논문은 가수 인식 작업을 통해 학습한 모델을 활용하여 다양한 음악적 특성을 반영하는 요소들 중 가수의 목소리 특성을 추출하는 방법을 제안한다. 배경음이 포함된 음원 역시 활용할 수 있지만, 음원에 포함된 배경음은 네트워크가 가수의 목소리를 온전하게 인식하는 것을 방해할 수 있다. 이를 해결하기 위해 본 연구에서는 음원 분리를 통해 배경음을 분리하는 사전 작업을 수행하고자 하며, SiSEC에 등장해 검증된 모델 구조를 활용하여 분리된 보컬로 이루어진 데이터 세트를 생성한다. 최종적으로 분리된 보컬을 활용하여 아티스트의 목소리를 반영하는 음색 기반 음악 특성을 발굴하고자 하며, 배경음이 분리되지 않은 음원을 활용한 기존 방법과의 비교를 통해 음원 분리의 효과를 알아보고자 한다.

Abstract

Preference for music is determined by a variety of factors, and identifying characteristics that reflect specific factors is important for music recommendations. In this paper, we propose a method to extract the singing voice related music features reflecting various musical characteristics by using a model learned for singer identification. The model can be trained using a music source containing a background accompaniment, but it may provide degraded singer identification performance. In order to mitigate this problem, this study performs a preliminary work to separate the background accompaniment, and creates a data set composed of separated vocals by using the proven model structure that appeared in SiSEC, Signal Separation and Evaluation Campaign. Finally, we use the separated vocals to discover the singing voice related music features that reflect the singer's voice. We compare the effects of source separation against existing methods that use music source without source separation.

Keyword : Singer identification, Music recommendation, Music representation learning, Timbre-related music similarity, Deep learning

a) SK텔레콤(SK Telecom)
‡ Corresponding Author : 이승진(Seungjin Lee)
E-mail: lee.seungjin@sktelecom.com
Tel: +82-31-710-5114
ORCID: https://orcid.org/0000-0001-7916-1947
· Manuscript received September 18, 2019; Revised November 18, 2019; Accepted November 26, 2019.

Ⅰ. 서 론

기술의 발달로 인하여 수많은 음악 콘텐츠가 공급될 수 있는 환경이 조성되었으며, 이는 콘텐츠 소비자들에게 선택의 어려움을 야기했다. 이러한 배경에서 다양한 연구가 진행되고 있으며, 음악 추천 및 음악 정보 검색 기술은 음악 스트리밍 서비스 및 인공지능 스피커의 등장과 맞물려 최근 주목을 받고 있는 분야다. 현재 음악 스트리밍 서비스에선 사용자 청취 이력을 활용한 개인화 추천, 전문가가 직접 제작한 플레이리스트 등을 활용하여 사용자에게 질 높은 추천 서비스를 제공하고 있다. 하지만, 전자의 경우 신곡에 대하여 추천이 불가능하고 추천의 결과가 편향될 수 있으며, 후자의 경우 전문가의 노동이 지속적으로 필요하다. 이러한 배경에서 음원 신호 자체를 활용하는 방법에 대한 수요는 높아지고 있다.

기술의 발달과 함께 전통적인 특성 발굴의 대안으로 특성 학습이 등장하면서, 신경망 학습을 활용하여 음원 신호 기반 유사한 음악을 찾아주는 기술이 주목을 받고 있다. 전통적으로는 도메인 지식을 활용하여 특성을 추출하고, 다양한 기법을 활용하여 유사한 음악을 분석하는 것이 일반적이었다. 대표적으로는 MFCC(Mel Frequency Cepstral Coefficients)와 K-means 군집화를 활용하여 음악 간의 유사성을 계산하는 연구 [1]이 있었으며 감정 인식, 언어 인식, 오디오 인식 등에서 성공적으로 사용되었던 발화 기반 특성인 I-Vector를 추출한 후 음색기반 음악 간의 유사성을 계산하고 아티스트 인식 작업을 수행한 연구 [2]가 있었다. 이미지 영역에서 심층 합성곱 신경망을 기반으로 하는 분석 기술이 비약적인 발전을 이룬 후, 음원 신호 영역에도 시간-주파수 특성을 이미지화 한 딥 러닝 기반의 분석 기술이 널리 적용되기 시작했다. 대표적인 특성 학습 연구로는 아티스트 정보를 학습하여 특성 벡터를 추출한 후 음악 간의 유사성을 계산하고 이를 활용하여 장르 등을 분류하여 아티스트 정보기반 특성을 검증한 연구 [3]이 있다. 아티스트 정보를 활용하는 특성 학습은 장르, 성별, 악기, 분위기 등을 복합적으로 담고 있어 효과적인 지도 학습 방법이지만, 음악 간의 유사함을 결정하는 다양한 요소들 중 특정 요소를 명확하게 반영하지 못한다. 예를 들어, 사후적으로 해석하지 않으면 사용된 악기, 노래의 음색, 곡의 분위기 중 어떤 요소가 유사한지 알 수 없다. 최근 다양한 요소들 중 가수 인식 작업을 통해 음색기반 특성을 추출하고자 했던 연구 [4]가 있었으며, 보컬 신호를 활용하여 특성의 효과성을 입증하였다. 위 연구와 유사하게 가수 인식 학습에 녹음된 보컬 음원와 오리지널 음원을 활용하는 연구 [5]가 있다. 특성 벡터를 시각화 한 결과, 거리 기반 학습을 통해 동일한 가수의 배경음이 포함된 음원과 포함되지 않은 보컬 음원을 가까운 공간에 표현할 수 있음을 보였다. 그러나 수백만 개 이상의 곡들 사이에서 유사한 음악을 찾아야 하는 실제 서비스에서 모든 곡에 대한 녹음된 보컬 파일은 존재하지 않기에 위 논문을 실험 데이터가 아닌 실제 데이터로 학습한 모델을 직접 서비스에 활용하는 것은 불가능에 가깝다. 이를 해결하기 위한 방법으로 음원 분리를 통해 추출한 보컬 신호로 녹음된 보컬 파일을 대체하는 방법이 있다. 딥 러닝 기술의 발전으로 음원 분리 분야 역시 비약적인 성능의 개선이 있었으며, SiSEC(Signal Separation and Evaluation Campaign)이라는 음원 분리 경진 대회에서 DenseNet(densely connected convolutional network)과 LSTM(Long Short Term-Memory)을 결합한 MMDense-LSTM[6], skip-connection이 포함된 U-Net[7], Wave-U-Net[8] 등이 제출되었고 MMDenseLSTM이 최고 성능을 달성하였다[9].

본 논문은 음원 분리를 어플리케이션 레벨에서 활용하여 가수 인식을 위한 데이터셋을 생성하고, 가수 인식에 있어 음원 분리의 효과성 입증이 주된 연구 목표이다. 본 연구에서는 가수의 목소리를 반영하는 특성을 추출하기 위하여 크게 2가지 작업을 진행한다. 우선 전처리 작업을 통해 특성 학습을 위한 입력 데이터를 구성한다. 다음으로 가수 정보 기반 특성 학습을 통해 음원 분리 및 가창 등장구간 추출의 효과를 알아본다. MSD-singer[5] 데이터 세트 및 자체 생성한 한국 가요 데이터 세트에 대해 가수 인식 작업을 수행하여 음색기반 음악 특성을 추출하고, 비교 실험을 통해 오리지널 음원과 비교하고자 한다.

본 논문의 구성은 다음과 같다. 우선 II 장에서는 제안 방법의 전체적인 작업 순서 및 구조, 실험에 사용한 데이터 세트, 전처리, 모델 구조에 관하여 서술한다. 다음 III 장에서는 비교 실험을 통해 제안하는 방법이 기존 방법에 비해 높은 가수 인식 성능을 보임을 보이고, 손실 함수에 따른 실험 결

과를 분석한다. IV 장에서는 시각화 결과 및 정량적인 유사성 분석 결과를 서술한다. 마지막으로 V 장에서는 본 논문에서 진행한 연구 내용에 대한 요약 및 결론을 기술한다.

II. 제안 방법

II 장에서는 제안하는 방법의 전체적인 작업 순서 및 구조에 대해 간단하게 설명하고 실험에 사용되는 데이터 세트를 소개한다. 다음으로 세 가지 전처리 과정에 대하여 설명하고, 가수 인식 네트워크 구조에 대해 설명한다.

1. 전체 구조

제안하는 방법의 전체 구조는 그림 1과 같으며 전처리, 학습 및 추론 부분으로 나뉜다. 전처리 과정은 3단계로 우선 음원 분리를 통하여 분리된 보컬 신호를 추출한다. 다음으로 추출된 보컬 신호를 STFT(Short Time Fourier Transform)와 mel-filter를 통해 2차원 mel-spectrogram 이미지로 변환한다. 마지막으로 사전학습된 보컬 등장구간 추출 모델을 활용하여 실제 보컬이 등장하는 mel-spectrogram만 추출하는 것으로 전처리는 끝난다. 전처리를 통해 추출한 mel-spectrogram을 입력으로 가수 인식 작업 기반의 특징 학습을 진행하며, 최종적으로 새로운 음원이 들어왔을 때 학습된 가수 인식 모델로부터 음색기반 특성을 추출한다.

2. 데이터 세트

본 연구에서는 가수 인식 실험의 평가를 위해 두 가지 데이터 세트를 사용하였다.

- MSD-singer[5]: MSD(Million Song Dataset) 데이터에 가창 등장구간 추출을 적용하여 선행연구에서 제

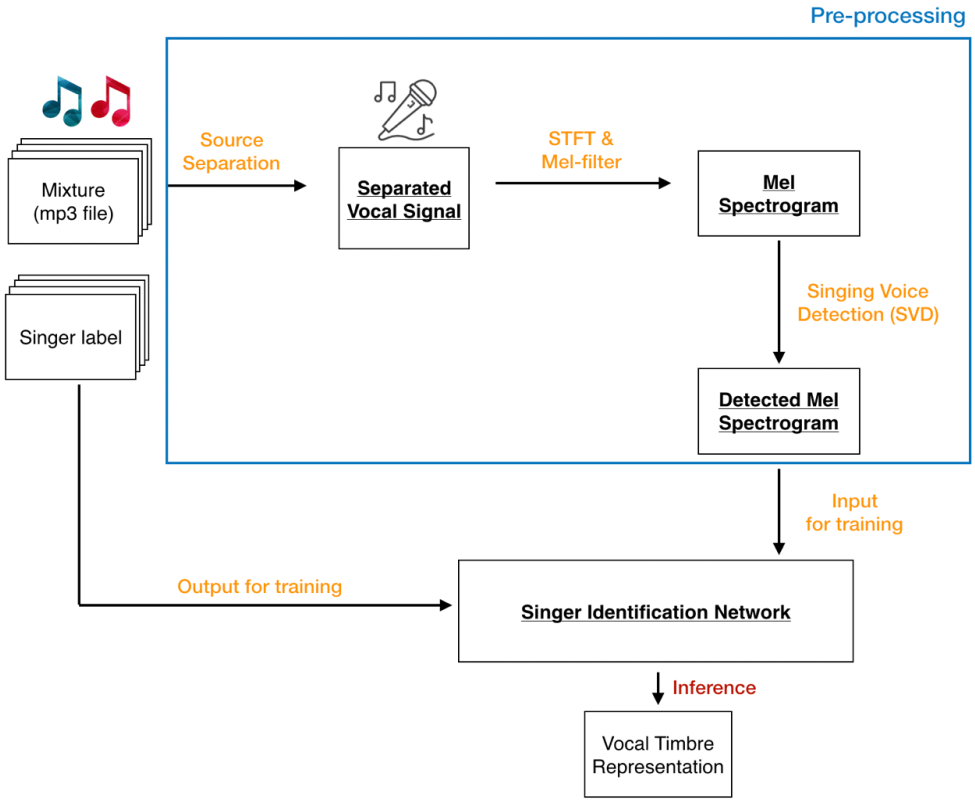

그림 1. 음색기반 음악 특성 추출에 관한 전체 개요도
Fig. 1. Overall flow diagram of timbre-related music feature extraction

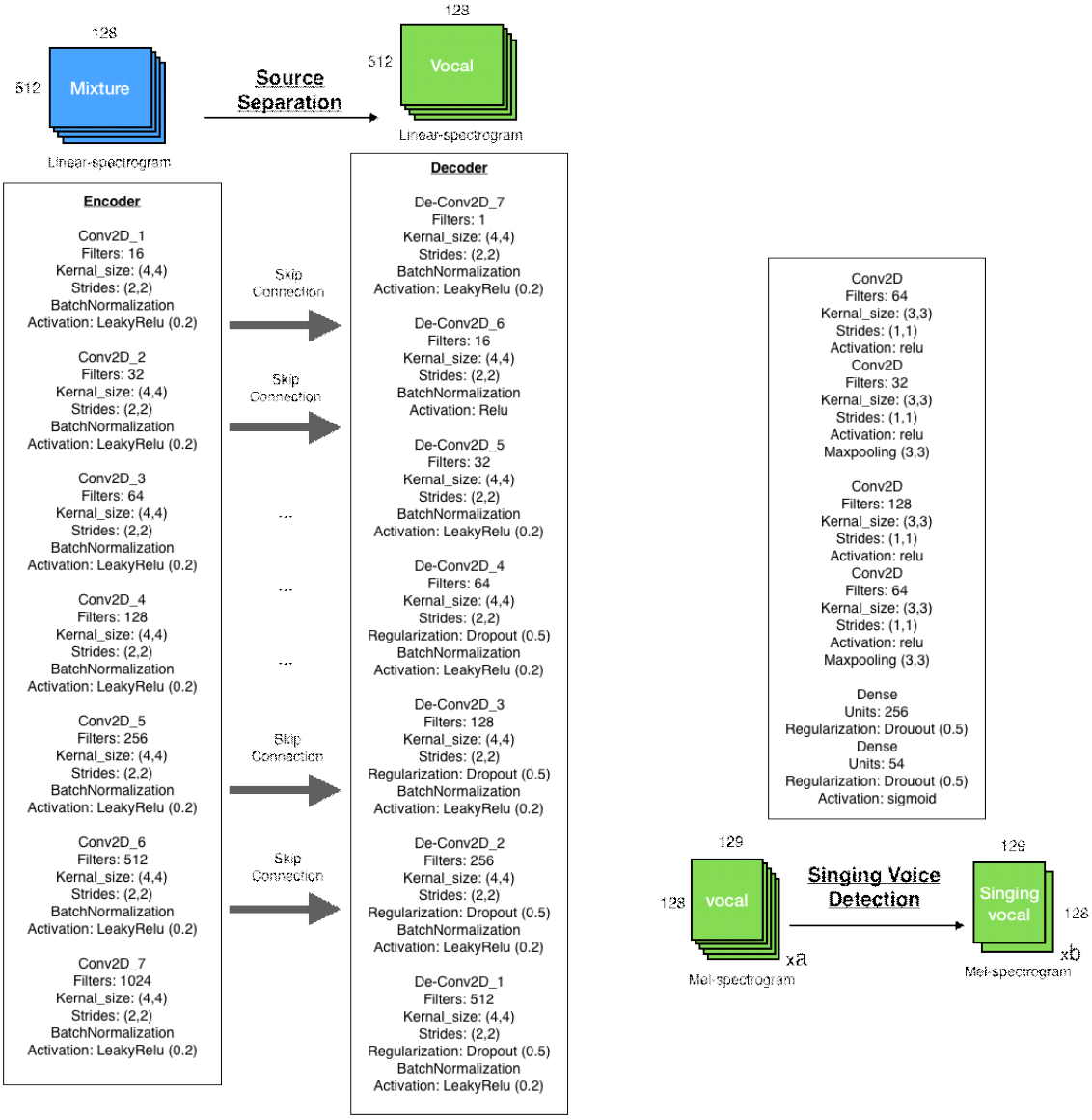

그림 2. 보컬 분리 모델 및 보컬 등장구간 추출 모델 상세
Fig. 2. Details of vocal source separation model and singing voice detection model

작해 사용한 데이터 세트로, 학습 데이터(20,000 songs, 1,000 singers)와 평가 데이터(10,000 songs, 500 singers)로 구성된다.

- 한국 가요(K-POP): 한국 가요(K-POP) 음원 및 이에 대응되는 보컬 음원을 분리하여 자체 제작한 데이터 세트로, 가창 등장구간 추출과 보컬 분리의 효과 비교를 위해 추가하였다. 학습 데이터(10,000 songs, 1,000 singers), 정확도 평가를 위한 평가 데이터1(5,000 songs, 500 singers), 성별 및 발매국 정보를 예측하기 위한 평가 데이터2(2,000 songs, 1,000 singers) 로 구성된다.

3. 전처리

본 연구에서의 전처리 과정은 세 가지로, 분리된 보컬을

추출하는 음원 분리 작업, 분리된 보컬 음원에서 mel-spectrogram을 추출하는 작업, 실제 보컬이 등장하는 프레임을 선별하는 작업을 포함한다.

음원 분리 학습 모델을 만들기 위하여 MUSDB18[10] 데이터 세트를 사용하였고, 학습 모델로는 U-Net 기반의 음원 분리 모델[7]을 사용하였다. 선행 연구에서 8,192 Hz로 다운 샘플링을 진행한 것과 달리, 본 연구에서는 분리된 보컬 신호 음원 데이터베이스를 확보하기 위하여 44,100Hz로 샘플링을 진행하였다. STFT의 fft size는 1,024, hop size는 256으로 설정하였다. 또한 보컬이 분리된 크기 정보와 사전에 추출했던 위상 정보를 활용하여 모든 곡에 대한 분리된 보컬 음원을 저장하였다. 음원 분리 작업의 상세 모델 구조는 그림 2의 왼쪽과 같다. skip-connection의 경우, encoder의 Conv2D_1과 decoder의 De-Conv2D_6, encoder의 Conv2D_2과 decoder의 De-Conv2D_5 처럼 블록 번호의 합이 7이 되는 레이어끼리 연결된다. encoder의 출력은 decoder의 입력이기에 연결이 존재하지않는다. decoder의 출력은 사전에 추출했던 위상 정보와 결합되어 분리된 보컬 신호 음원을 생성한다. 손실 함수 및 optimizer는 mean absolute error와 adam을 사용하였고 batch normalization을 적용하였다. 학습 파라미터로 batch size는 16을 적용하였고 learning rate는 0.0001을 적용하였다. 오디오 처리는 librosa[15], 신경망 모델 학습은 tensorflow[16]를 이용해 진행하였다.

다음으로 추출한 분리된 보컬 음원에 대하여 22,050Hz로 리샘플링 및 정규화하고 로그 스케일의 mel-spectrogram을 추출하였다. 이때, 세부 파라미터들은 Lee[5]의 연구와 동일하게 1,024의 fft size, 512의 hop size, 128의 mel band 로 설정하였다. 이렇게 추출된 mel-spectrogram은 약 3초 단위로 나뉘어 사용되며 (129, 128) 차원을 가진다. 모든 오디오 처리는 위와 마찬가지로 librosa[15]를 이용해 진행하였다.

전처리 마지막 작업으로 분리된 보컬 음원에서 실제 보컬이 등장하는 프레임을 선별하는 작업을 위해, MedleyDB[11] 데이터 세트를 사용하였다. MedleyDB 데이터는 정확한 보컬 위치에 대한 주석정보가 없었기에 Lee[13,14]의 방법과 동일하게 보컬 위치에 대한 정보를 생성하였다. CNN 기반의 모델[12]을 사용하여 약 3초 단위의 mel-spectrogram에 대해 보컬 여부를 이진분류한다. 보컬 위치 추정 작업의 상세한 모델은 그림 2의 오른쪽에 해당한다. 입력 데이터는 전체 곡을 약 3초 단위로 a개 추출하여 (a, 129, 128)의 차원을 가지며, 보컬로 분류된 b개의 mel-spectrogram을 추출한다. 예를 들어, 보컬이 등장하지 않는 음악의 경우 mel-spectrogram을 추출할 수 없기 때문에 이 경우 음색 기반 특성은 존재하지 않는다. 손실 함수 및 optimizer는 cross entropy와 adam을 사용하였다. 학습 파라미터로 batch size는 128을 적용하였고 learning rate는 0.001을 적용하였다. 오디오 처리는 librosa[15], 신경망 모델 학습은 tensorflow[16]를 이용해 진행하였다.

4. 가수 인식 모델

본 연구의 네트워크 구조는 앞서 언급한 선행 연구[1,5]의 기본 구조와 유사하다. 기본 모델은 5개의 1차원 합성곱 신경망 층으로 구성되며, 그림 3와 같다. 처음 합성곱 신경망 층은 80개의 필터(size: 4, stride: 1)와 LeakyReLU로 구성되어 있으며 batch normalization과 max pooling(size: 4)을 적용하였다. 이어지는 3개의 합성곱 신경망 층은 필터의 수가 128개인 것을 제외하면 첫번째 층과 동일하게 구성된다. 마지막 합성곱 신경망 층은 256차원의 특성을 추출할 수 있도록 256개의 필터로 구성되어 있으며 과적합을 막기위해 50%의 Dropout이 함께 적용된다. 입력 값으로는, 앞서 추출한 약 3초 단위로 분할된 mel-spectrogram이 사용된다. 이때, 모든 합성곱 작업은 시간을 나타내는 차원에서만 수행된다.

본 연구에서는 categorical cross entropy와 hinge loss 두 가지로 손실 함수를 달리해 가수 인식 모델을 학습하였다. 이후 등장할 표에서의 설명을 위해 categorical cross entropy는 CCE, hinge loss는 HL로 표시한다. 그림 4 구조는 분류 문제로 가수 인식 모델을 학습시키는 방법이며, 식 (1)과 같은 손실 함수를 사용하여 학습에 사용된 가수의 수가 2,000이라면 2,000명의 가수 중 한 명의 가수를 식별한다.

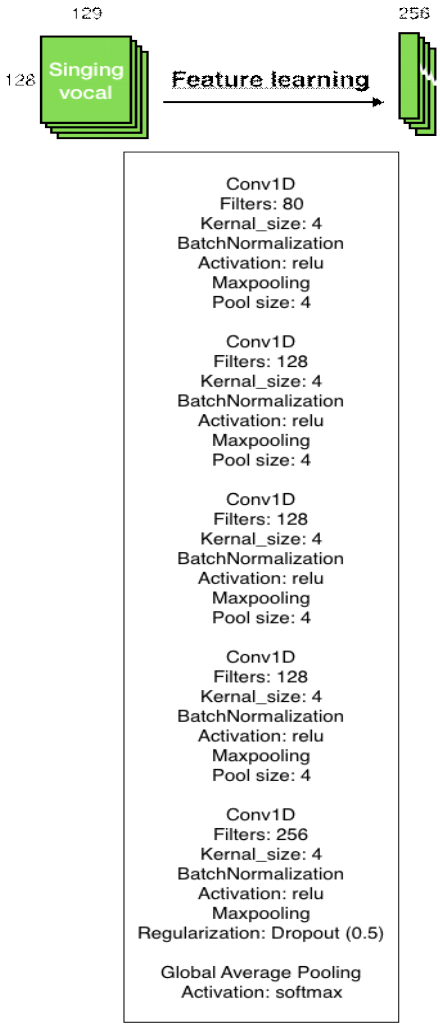

그림 3. 가수 인식 모델 상세
Fig. 3. Details of singer identififcation model

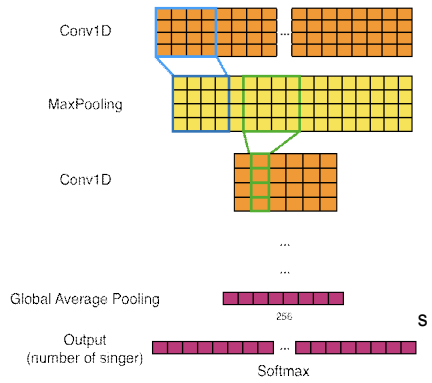

그림 4. 가수 인식 네트워크 구조 (분류)
Fig. 4. Singer identification network structure (classification)

$$categorical\ cross\ entropy = -\frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{J} y_j \cdot \log(\hat{y}_j) + (1-y_j)\cdot \log(1-\hat{y}_j) \quad (1)$$

위 방법은 직관적이지만, 학습에 사용되는 가수의 수가 매우 큰 경우 학습에 매개변수의 수가 매우 커져 모델이 무거워지는 단점이 존재한다. 그림 5 구조는 이러한 단점을 극복할 수 있는 모델이다. 거리 학습은 범주가 달라 서로 다른 특징으로 분류해야만 하지만, 전체적인 특성이 유사하여 같은 공간 내 있는 데이터를 떨어뜨리는 효과가 있다. 학습 단계에서 쿼리 노래와 같은 가수의 노래의 거리는 가깝게, 쿼리 노래와 다른 가수의 노래는 멀게 학습되며, 이를 통해 목소리가 비슷한 다른 가수의 특성 벡터를 잘 표현할 수 있다. 거리는 마지막 합성곱 층을 거친 256차원의 특성 간의 코사인 유사도(cosine similarity)를 기반으로 계산되며, 계산은 식 (2)과 같이 한다.

$$\cos(A,B) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \quad (2)$$

가깝게 만들 같은 가수의 노래를 positive sample, 멀어지게 만들 다른 가수의 노래를 negative sample로 한다. 여러 개의 negative sample을 추출하는 것 역시 가능하며 본 연구에서는 1부터 10까지의 수 중 가수인식 정확도가 가장 높았던 5를 negative sample로 설정하였다. 손실 함수는 식 (3)와 같다.

$$hinge\ loss(A,B) = \max\left[0, m - D(A,B_{positive}) + D(A,B_{positive})\right] \quad (3)$$

위 식에서 m은 margin을 나타내며 선행 연구[5]와 동일하게 0.1로 설정하였다. 학습 관련하여 미니 배치(mini-batch)의 크기는 32, 학습률(learning rate)는 0.001로 설정하였으며 과적합을 막기 위하여 10 epoch 동안 검증 데이터에 대한 손실(validation loss)의 개선이 없다면 학습을 멈추도록 설계하였다. 전처리 과정과 마찬가지로 모든 오디오 처리는 librosa[15], 신경망 모델 학습은 tensorflow[16]를 이용해 진행하였다.

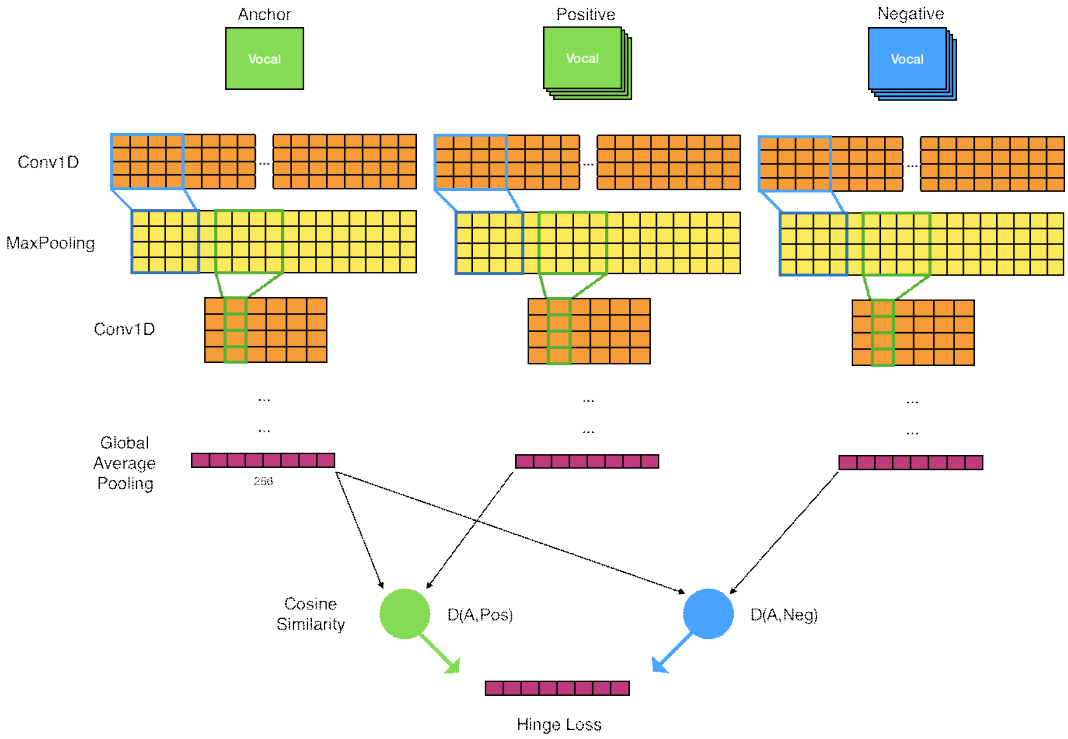

그림 5. 가수 인식 네트워크 구조 (거리 기반 학습)
Fig. 5. Singer identification network structure (metric learning)

표 1. 손실 함수에 따른 모델 파라미터 수
Table 1. The number of model parameters according to loss function

Model parameter	Categorical cross entropy (CCE)	Hinge loss (HL)
Trainable parameter	488,680	231,536
Non-trainable parameter	1,536	1,536
Total parameter	490,216	233,216
Optimizer	Adam	SGD
Learning rate	0.001	0.01
Batch size	256	32

III . 실험 결과

III 장에서는 오리지널 음원 기반 방법과 보컬 분리 기반 방법 간의 비교 실험을 통해 보컬 분리의 효과를 살펴보고자 한다. 또한 한국 가요로 구성된 데이터를 활용해 보컬 구간 추정 효과 및 보컬 분리 효과에 대해 살펴보고자 한다.

1. 실험 방법 및 평가 지표

본 실험에서는 배경음이 포함된 신호를 사용한 방법과 제안하는 방법의 비교 및 손실 함수 간의 정량적인 비교 실험 결과를 서술한다. 우선 음원 분리 여부와 손실 함수 변경에 따른 가수 인식 성능을 평가하기 위해 MSD-singer 데이터의 학습에 사용되지 않은 500명의 가수를 활용하여 평가를 진행한다. 우선 각 가수 별 전체 20곡 중 15곡을 사용하여 가수 인식 모델을 만든다. 만들어진 모델에서의 평가를 위하여 나머지 5곡을 평가를 위한 쿼리곡으로 활용하며 총 2,500개의 쿼리를 생성한다. 코사인 유사도를 기준으로 가까운 가수를 검색하였으며 쿼리와 거리가 가장 가까운 가수가 선택되는 방식으로 정확도를 측정한다. 정확도 관련 평가 지표로는 거리가 가장 가까운 가수가 쿼리 가수와 일치하는지를 측정하는 top1 정확도(top1 accuracy), 거리가 가까운 가수 5위 안에 쿼리 가수가 포함되는지를 측정하는 top5 정확도(top5 accuracy)를 사용한다. 또한 조금 더

추천에 영향을 미치는 세부적인 가수 인식 성능을 확인하기 위해서 정밀도(precision), 재현율(recall)을 사용하였다. Precision@k는 k개의 상위 항목들 중 가수가 일치하는 비율을 나타내며, recall@k는 평가에 사용된 전체 가수들 중 k개의 가까운 상위 항목에서 가수가 일치하는 비율을 나타낸다. 예를 들어 10개의 상위 항목들 중 5곡이 일치했다면, precision@10은 10곡 중 5곡이 일치하여 0.5, recall@10은 15곡중 5곡이 일치하여 0.333이 된다. 추가로 추천의 순서를 고려하는 평가를 위해 평균 정밀도를 다시 한번 더 평균낸 mAP(mean average precision)를 사용한다.

다음으로 한국 가요로 구성된 데이터 세트를 활용하여 보컬 구간 추정 모델 적용에 따른 효과와의 차이를 살펴보고자 한다. 평가를 위한 데이터로는 정확도 평가를 위한 평가 데이터1(5,000 songs, 500 singers) , 성별 및 발매국 정보를 예측하기 위한 평가 데이터2(2,000 songs, 1,000 singers) 를 활용하여 평가를 진행한다. 평가 지표로는 top1 정확도, Recall@5, Recall@10을 사용며, 가수의 성별, 국적을 정답으로 두고 mAP를 측정하여 모델 별 가수의 성별, 국적을 반영하는 정도의 차이 역시 확인하고자 한다.

2. 실험 결과

표 2는 음원 분리 여부 및 손실 함수 변경에 따른 가수 인식 정확도를 나타낸다. 의도했던 대로 나머지 조건이 동일하다면 분리된 보컬 기반 모델이 오리지널 음원 기반 모델보다 높은 정확도를 보임을 확인하였다. 이는 동일한 가수의 곡들 사이에서도 악기 구성, 장르 등이 달라지면 배경음의 차이가 발생할 수 있어 가수 인식 정확도가 떨어질 수 있음을 보여준다. 표 2에서 역시 분리된 보컬 기반 모델이 오리지널 음원 기반 모델보다 높은 성능을 보였으며, 추천의 순서를 반영하는 평가 지표인 mAP에서 역시 유의미한 차이를 보였다. mAP가 높은 경우 상위 추천 결과에 정답이 많이 분포하며, 이는 유사 음색 추천 관점에서의 분리된 보컬 기반 모델의 성능이 오리지널 음원 기반 모델의 성능과 비교해서 좋은 성능을 보임을 나타낸다. 가수에서 그룹 단위의 아티스트로 확장한다면 위와 같은 배경음의 차이가 커질 가능성이 더 커지기에, 음색을 반영하는 측면에서는 두 모델의 성능 차이는 더 커질 수 있을 것으로 예상된다.

손실 함수의 차이에 따른 성능 차이는 존재했으며, 표 2, 표 3의 결과 모두 거리 기반 학습 모델보다 분류 학습을

표 2. 음원 분리 여부 및 손실 함수 변경에 따른 가수 인식 정확도
Table 2. Singer identification accuracy according to the vocal separation and loss function

	Singer Identification Accuracy			
	CCE-MIXTURE	CCE-VOCAL	HL-MIXTURE	HL-VOCAL
Top 1 Accuracy	0.486	**0.508**	0.412	0.437
Top 5 Accuracy	0.716	**0.725**	0.672	0.704

표 3. 음원 분리 여부 및 손실 함수 변경에 따른 가수 검색 결과
Table 3. Singer retrieval result according to the vocal separation and loss function

	Singer Retrieval Results			
	CCE-MIXTURE	CCE-VOCAL	HL-MIXTURE	HL-VOCAL
mAP	0.254	**0.265**	0.221	0.233
Precision@5	0.341	**0.36**	0.3	0.314
Precision@10	0.257	**0.273**	0.242	0.25
Recall@5	0.113	**0.121**	0.1	0.106
Recall@10	0.171	**0.182**	0.161	0.167

진행했던 모델이 더 높은 성능을 보였다. 이는 학습 데이터의 곡 수가 20,000곡으로 비교적 많지 않았으며 학습 데이터를 구성하는 가수의 수 역시 충분하지 않아서 분류 모델보다 낮은 성능을 보였던 것으로 사료된다. 두 손실 함수 모델 간 정확도 차이를 비교했을 때, top1에서의 정확도 차이 보다 top5 정확도 차이가 적었으며, 표 2에서 역시 마찬가지의 결과를 보였다.

다음으로 한국 가요가 포함된 데이터 세트를 활용하여 보컬 등장구간 추정(SVD) 및 보컬 분리 적용에 따른 효과를 살펴보았으며 가수 인식 결과는 표 4과 같다. 비교 대상 모델은 보컬 등장구간 추정을 적용하지 않은 오리지널 음원 기반 모델, 보컬 등장구간 추정을 적용한 오리지널 음원 기반 모델, 보컬 등장구간 추정을 적용한 분리된 보컬 기반 모델 총 세 가지이다. 보컬 등장구간 추정 모델은 비교적 높은 정확도의 구간 추정을 위해 오리지널 음원이 아닌 분리된 음원 기반으로 학습을 진행하였다. 평가 데이터1(5,000 songs, 500 singers)을 활용하여 평가하였으며, 세 가지 모델 중 보컬 등장구간 추정을 적용한 분리된 보컬 기반 모델이 가장 높은 정확도를 보였으며 뒤이어 보컬 등장구간 추정을 적용한 오리지널 음원, 보컬 등장구간 추정을 적용하지 않은 오리지널 음원 순으로 정확도를 보였다. 보컬 등장구간 추정을 적용한 경우와 비교하여 음원 분리를 적용한 경우의 성능 향상폭이 2배 이상 컸으며, 이는 음원 분리의 효과를 입증하는 결과이다.

표 4. 보컬 등장구간 추정 및 보컬 분리 적용에 따른 가수 인식 결과
Table 4. Singer identification results according to the singing voice detection(SVD) and vocal separation

K-POP DATASET	Singer Identification Accuracy		
	MIXTURE (without SVD)	MIXTURE (SVD)	VOCAL (SVD)
Top 1 Accuracy	0.407	0.464	**0.613**
Top 5 Accuracy	0.648	0.708	**0.821**

표 5는 보컬 등장구간 추정 및 보컬 분리 적용에 따른 가수 성별 검색 결과를 나타낸다. 평가 데이터2(2,000 songs, 1,000 singers)를 활용하여 평가하였으며, mAP 측정을 통하여 높은 순위에 등장하는 가수의 성별이 쿼리 가수의 성별과 얼마나 동일한지를 알아보고자 하였다. 위 결과와 마찬가지로 세 가지 모델 중 보컬 등장구간 추정과 음원 분리를 모두 적용한 모델이 가장 높은 성능을 보였다. 하지만, 보컬 등장구간 추정과 음원 분리의 효과를 비교했을 때, 보컬 등장구간의 효과가 더 컸다. 이는 성별을 인식함에 있어서는 보컬을 분리하는 것만큼 보컬 등장구간을 찾아주는 것 역시 중요하다는 점을 보여준다.

표 5. 보컬 등장구간 추정 및 보컬 분리 적용에 따른 가수 성별 검색 결과
Table 5. Singer gender retrieval results according to the singing voice detection(SVD) and vocal separation

	Singer Gender Retrieval Results		
	MIXTURE (without SVD)	MIXTURE (SVD)	VOCAL (SVD)
mAP	0.808	0.87	0.89

표 6. 보컬 등장구간 추정 및 보컬 분리 적용에 따른 발매국 검색 결과
Table 6. Track release country retrieval results according to the singing voice detection(SVD) and vocal separation

	Track Release Country Retrieval Results		
	MIXTURE (without SVD)	MIXTURE (SVD)	VOCAL (SVD)
mAP	0.686	0.691	0.695

표 6은 보컬 등장구간 추정 및 보컬 분리 적용에 따른 곡의 발매국 검색 결과를 나타낸다. 발매국은 한국 가요와 그 외(MSD)로 이진화되었으며 성별과 마찬가지로 해당 특성이 추천에 있어 발매국 정보를 반영하는 효과가 있는지를 알아보고자 하였다. 제안하는 모델이 가장 높은 성능을 보였으나, 그 차이가 유의미하지 않았다. 이를 통해 구간 추정 및 보컬 분리 적용과 발매국 정보와는 큰 연관성이 없음을 확인할 수 있었다.

Ⅳ. 결 론

본 논문에서는 최근 인공지능 스피커의 등장과 맞물려 중요하게 떠오른 유사한 음악을 찾아주는 기술과 관련하여, 보컬 분리와 가수 인식 학습을 통해 음색기반 유사성을 정

의하는 특성을 발굴하고자 하였다. 이를 위해 가수 인식 작업을 통한 음원 분리의 효과성을 탐색하였고 특성 추출 순서 및 방법을 제안하였다.

실험을 위한 데이터 세트는 Lee [5]의 MSD-singer 데이터 세트와 한국 가요 데이터 세트를 가공하여 사용하였으며, 각 모델에 대하여 top1 정확도, top5 정확도, precision@5, precision@10, recall@5, recall@10, mAP 측면에서 평가하였다. 제안하는 분리된 보컬 기반 방법은 오리지널 음원 기반 방법 및 보컬 등장구간 추정을 적용한 모델에 비해 모든 면에서 더 높은 가수 인식 성능을 보였으며, 음색 기반 유사 음악 및 가수 추천 서비스에 응용될 수 있는 가능성을 보였다.

향후 연구로는 우선 기학습된 보컬 등장구간 추정 모델과 보컬 분리 모델을 통해 보다 많은 음원에 대한 데이터를 확보하고자 한다. Park [1]의 연구를 비롯한 많은 연구들에서 성능이 검증된 거리 학습 기반 방법의 효과를 방대한 데이터 세트를 활용해 재실험하고자 한다. 다음으로는 multi-task learning을 통해 분위기, 스타일 등의 특성을 동시에 선택적으로 반영할 수 있는 음악 추천을 위한 특성을 발굴하고, 사용자 프로파일링과의 접목을 통해 본 연구를 개인화 추천으로 확장시키고자 한다.

참 고 문 헌 (References)

[1] J. Park, J. Lee, J. Park, J. Ha, J. Nam, "Representation Learning of Music Using Artist Labels", Proceeding of International Society for Music Information Retrieval Conference, Paris, France, pp. 717-724, 2018.

[2] B. Logan, A. Salomon, "A Music Similarity Function Based on Signal Analysis", ICME, Tokyo, Japen, pp. 22-25, 2001.

[3] H. Eghbal-Zadeh, B. Lehner, M. Schedl, G. Widmer, "I-Vectors for Timbre-Based Music Similarity and Music Artist Classification", Proceeding of International Society for Music Information Retrieval Conference, Malaga ,Spain pp. 554-560, 2015.

[4] C. I. Wang, G. Tzanetakis, "Singing style investigation by residual siamese convolutional neural networks", Proceeding of International Conference Acoustic, Speech and Signal Processing, Calgary, Canada, pp. 116-120, 2018.

[5] K. Lee, J. Nam, "LEARNING A JOINT EMBEDDING SPACE OF MONOPHONIC AND MIXED MUSIC SIGNALS FOR SINGING VOICE", Proceeding of International Society for Music Information Retrieval Conference, Delft, Netherlands, 2019.

[6] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, Y. Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending", Proceeding of International Conference Acoustic, Speech and Signal Processing, New Orleans, LA, USA, pp. 261-265, 2017.

[7] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, T. Weyde, "Singing voice separation with deep U-Net convolutional networks", Proceeding of International Society for Music Information Retrieval Conference, Suzhou, China, pp. 745-751, 2017.

[8] D. Stoller, S. Ewert, S. Dixon, "Wave-u-net: A multi- scale neural network for end-to-end source separation", Proceeding of International Society for Music Information Retrieval Conference, Paris, France, pp. 334-340, 2018.

[9] D. Ward, R. D. Mason, C. Kim, F. R. Stoter, A. Liutkus, M. Plumbley, "SISEC 2018: state of the art in musical audio source separation-Subjective selection of the best algorithm", proceeding of the 4th Workshop on Intelligent Music Production, 2018.

[10] Z. Rafii, A. Liutkus, F. R. Stoter, S. I. Mimilakis, R. Bittener, "The MUSDB18 corpus for music separation", 2017 Zafar Rafii, Antoine Liutkus, Fabian Rovert-Stoter, Stylianos loannis Mimiiakis, Rachel Bittner. MUSDB18 - a corpus for music separation, 2017, <10.5281/zenodo.1117371>. <hal-02190845>

[11] R. Bittener, J. Salamon, M. Tierney, M. Mauch, C. Cannam, J. P. Bello, "MedleyDB: A multitrack dataset for annotation-intensive mir research, Proceeding of International Society for Music Information Retrieval Conference, Taipei, Taiwan, pp. 155-160, 2014.

[12] J. Schluter, T. Grill, "Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks", Proceeding of International Society for Music Information Retrieval Conference, Malaga, Spain, pp. 121-126, 2015.

[13] K. Lee, K. Choi, J. Nam, "Revisiting Singing Voice Detection: a quantitative review and the future outlook", Proceeding of International Society for Music Information Retrieval Conference, Paris, France, pp. 506-513, 2018.

[14] J. Schluter, "Learning to pinpoint singing voice from weakly labeled examples", Proceeding of International Society for Music Information Retrieval Conference, New York, USA, pp. 44-50, 2016.

[15] B. Mcfee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, O. Neito, "Librosa: Audio and music signal analysis in python", Proceeding of the 14th Python in Science Conference, 2015.

[16] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Kudlur, "Tensorflow: a system for large-scale machine learning", Proceeding of the 12th USENIX conference on OSDI, 2016.

저 자 소 개

이 승 진

- 2019년 ~ 현재: SK텔레콤 ICT기술센터
- ORCID : https://orcid.org/0000-0001-7916-1947
- 주관심분야 : 추천 기술, 멀티미디어 분석 및 처리