너의 다음곡이 보여

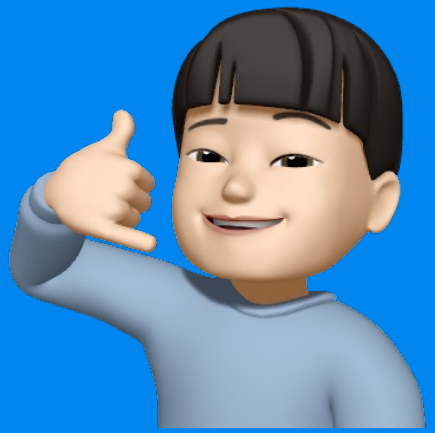
: Self Supervised learning을 이용한 음색 기반 노래 추천 시스템

이지평 장성현 김보현 김종윤 김정하

팀원 소개

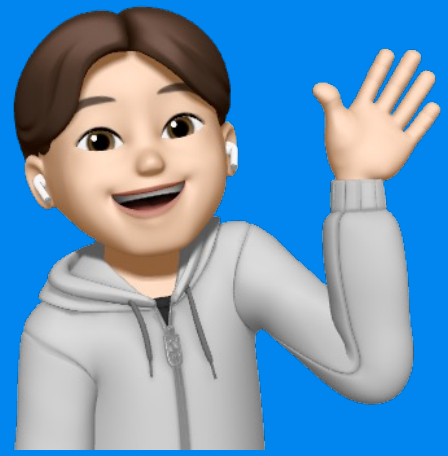
이지평

김정하

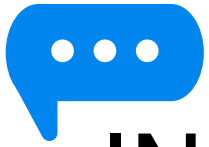


장성현

김보현



김종윤



INDEX

Previous work

- 이전 Framework
- 수정된 Framework

TCAE

- TCAE

Experiments

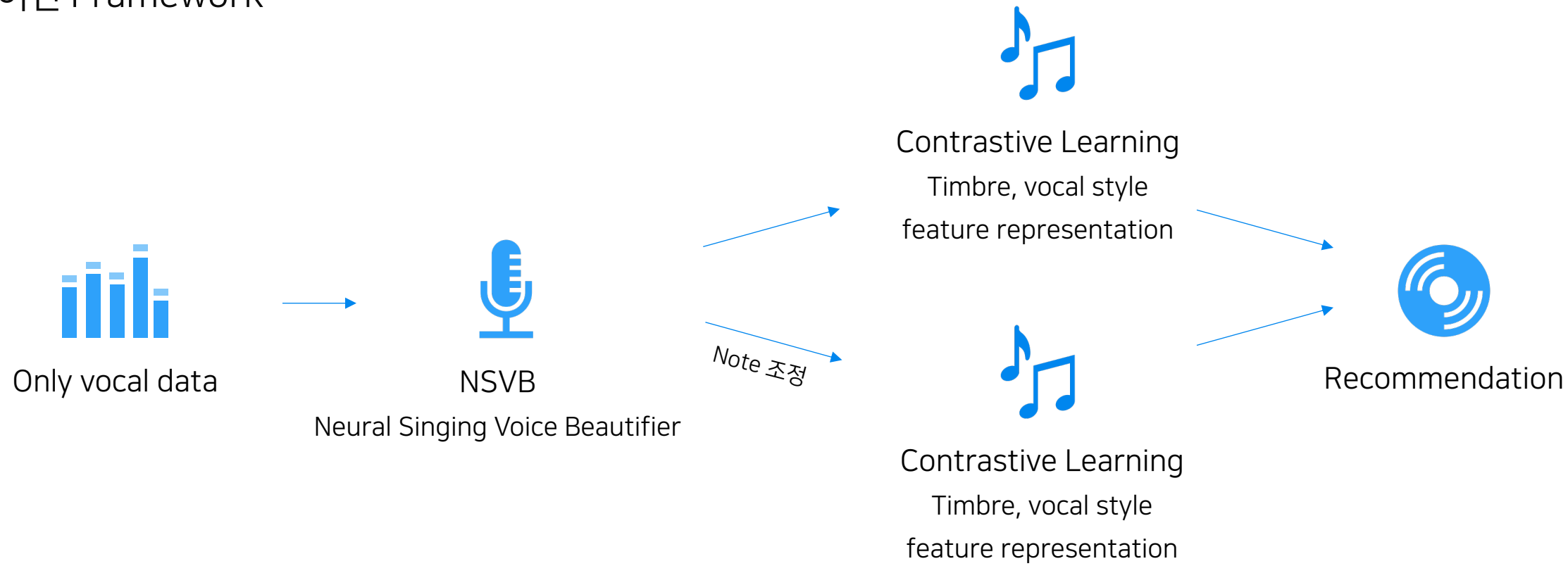
- 문제점
- 데이터 재구성
- 실험 결과

진행 계획

01

Previous work

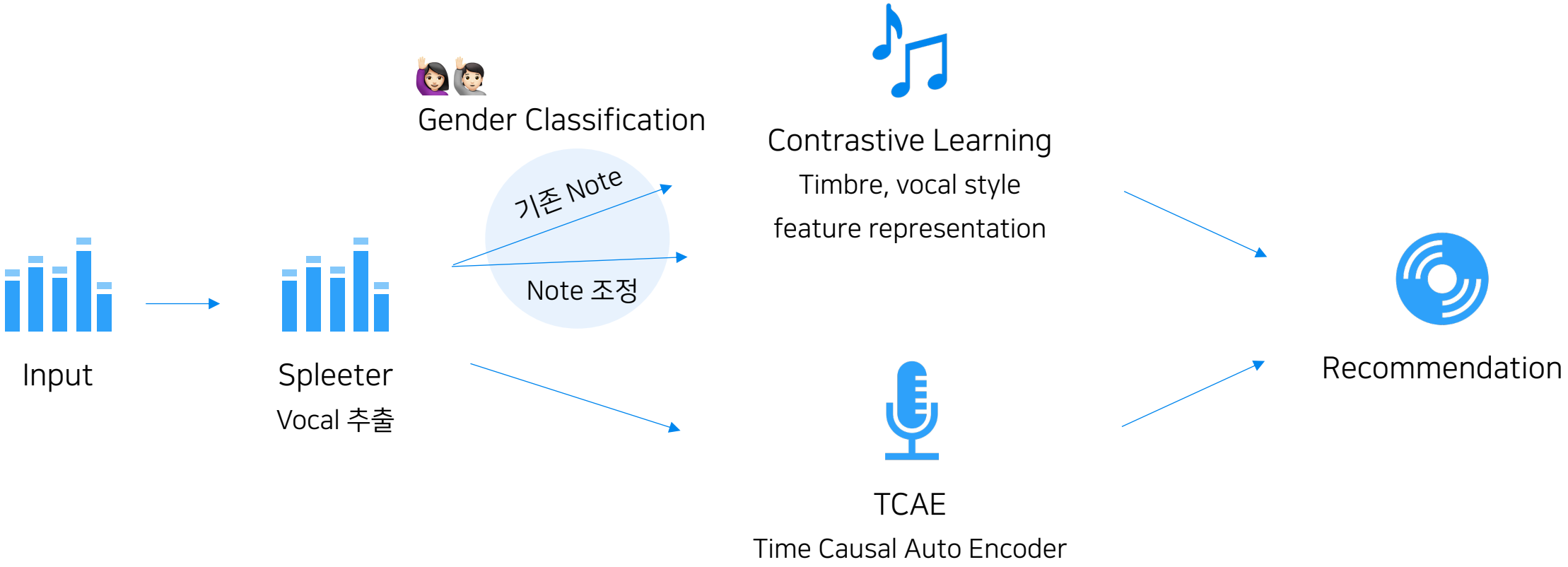
이전 Framework



01

Previous work

수정된 Framework



02

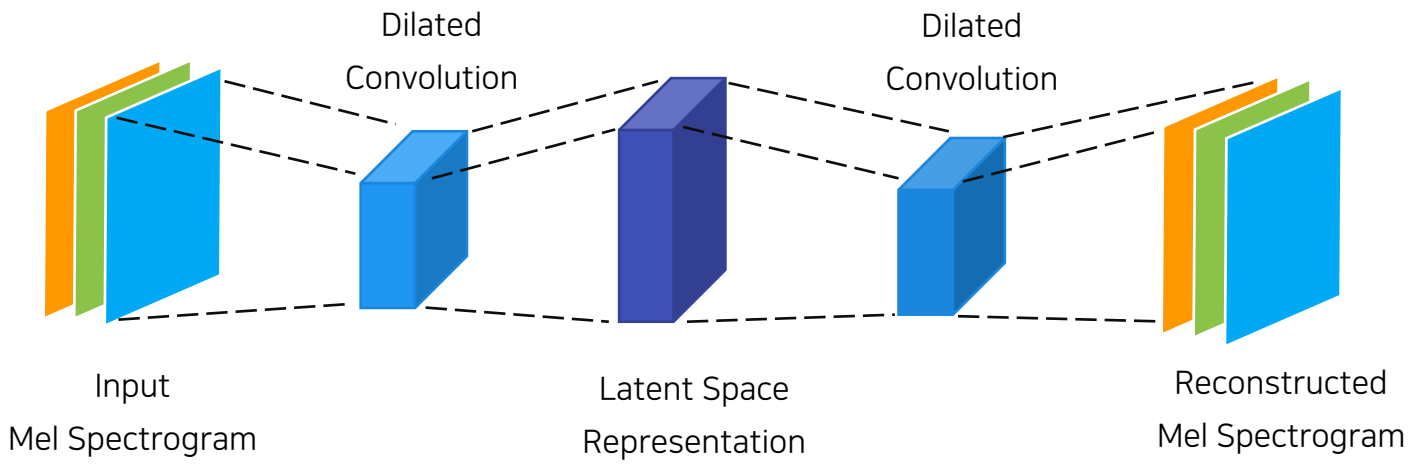
TCAE

Time Causal Auto Encoder

TCAE

Contrastive learning을 통해 반영할 수 있는 Timbre, vocal style과 더불어,

Vocal의 특징을 더욱 풍부하게 활용하고자 멜로디 정보를 반영할 수 있는 TCAE 사용



02

TCAE

Time Causal Auto Encoder

TCAE

- 데이터

Mel Spectrogram

Sampling rates: 16000, Window sizes: 512, Hop Lengths: 256, Mel Filters: 48

한 곡의 Mel Spectrogram shape: (48,7501)

- 구조

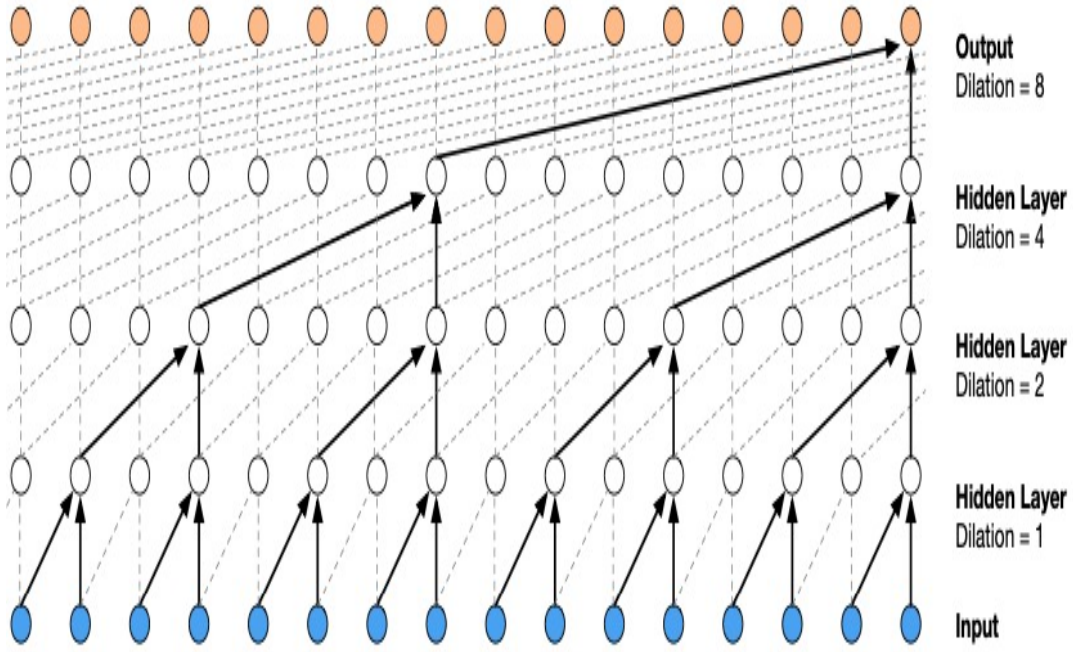
Dilated Causal Convolution Layers

AutoEncoder

02

TCAE

Dilated Causal Convolution Layers



Dilated convolution :

- 필터가 특정 단계에서 입력값을 건너뛰어 길이보다 큰 영역에 걸쳐 적용되는 convolution
- 일반 convolution layer보다 dilated convolution layer가 효과적으로 작동함
- stacked dilated convolution은 많지 않은 layer로 매우 큰 수용 필드를 가질 수 있게 하며, 입력값과 계산 효율성을 보존할 수 있게 함

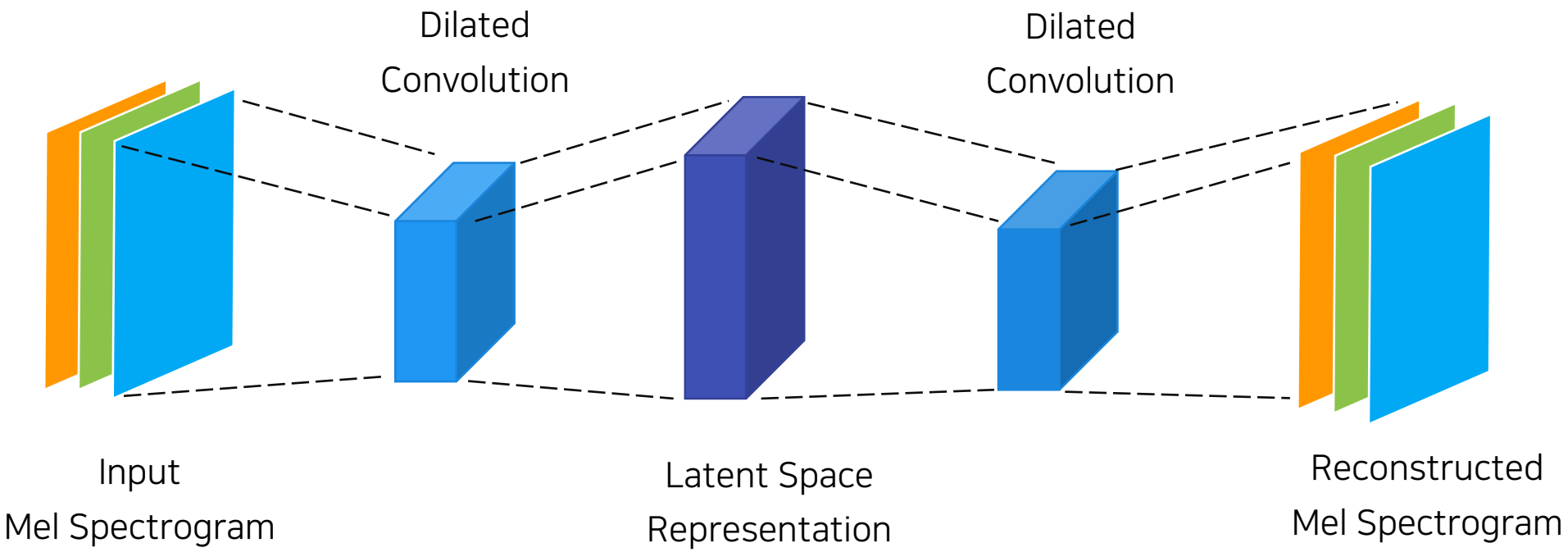
장점 :

- dilation factor를 증가시키면 receptive field의 성장이 일어남
- block을 쌓을 수록 크기가 더욱 커짐

02

TCAE

Model Structure



02

TCAE

요약

- Dilated Causal Convolution은 시간에 대해 큰 수용 필드를 사용함으로써 긴 시간의 흐름에 대해 학습 가능
- TCAE의 Auto Encoder를 시간 의존성을 반영할 수 있는 Dilated Causal Convolution Layers로 구성
- TCAE의 Latent Space는 **음악의 멜로디 정보를 담고 있는 임베딩**
- **Latent Space의 유사성은 비슷한 멜로디를 갖는다고 볼 수 있음**

03

Experiments

문제점

기존 데이터셋

30곡 이상 보유한 솔로 가수 150명의 음원 크롤링

30초 ~ 2분 30초 길이로 음원 자름

MR제거

- Contrastive learning에서 Loss가 수렴하지 않음
  특히, pitch shift를 negative pair,
  time stretch를 positive pair로 설정하고 학습했을 때 loss가 200

- TCAE에서 Input으로 들어가는 mel spectrogram의 길이가 다름
  음원의 총 길이가 2분 30초가 넘지 않는 곡 존재

03

Experiments

데이터 재구성

기존 데이터셋

30곡 이상 보유한 솔로 가수 150명의 음원 크롤링

30초 ~ 2분 30초 길이로 음원 자름

MR제거

- Contrastive learning에서 Loss가 수렴하지 않음
  특히, pitch shift를 negative pair,
  time stretch를 positive pair로 설정하고 학습했을 때 loss가 200

가수마다 음원의 개수가 다르고,
그 편차가 커서 학습이 잘 안된다고 판단

→ 한 가수당 30곡만 남기고 모두 삭제

03

Experiments

데이터 재구성

기존 데이터셋

30곡 이상 보유한 솔로 가수 150명의 음원 크롤링

30초 ~ 2분 30초 길이로 음원 자름

MR제거

- TCAE에서 Input으로 들어가는 mel spectrogram의 길이가 다름
  음원의 총 길이가 2분 30초가 넘지 않는 곡 존재

음원이 2분 30초가 넘지 않는 곡은 삭제 후
다른 음원으로 모두 대체

03

Experiments

실험 결과 : Contrastive learning

- 데이터 재구성 전 성능

Pitch Shift	pos	pos	neg	neg
Time Stretch	pos	neg	pos	neg
Loss	15	14	210	43

03

Experiments

실험 결과 : Contrastive learning

- 데이터 재구성 후 성능

Pitch Shift	pos	pos	neg	neg
Time Stretch	pos	neg	pos	neg
Loss	13	13	14	14

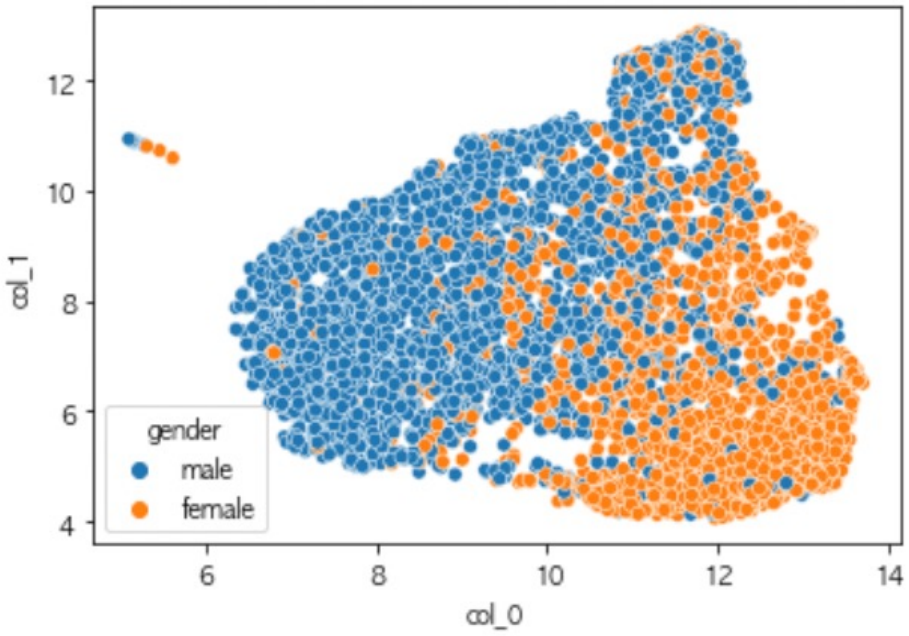
03

Experiments

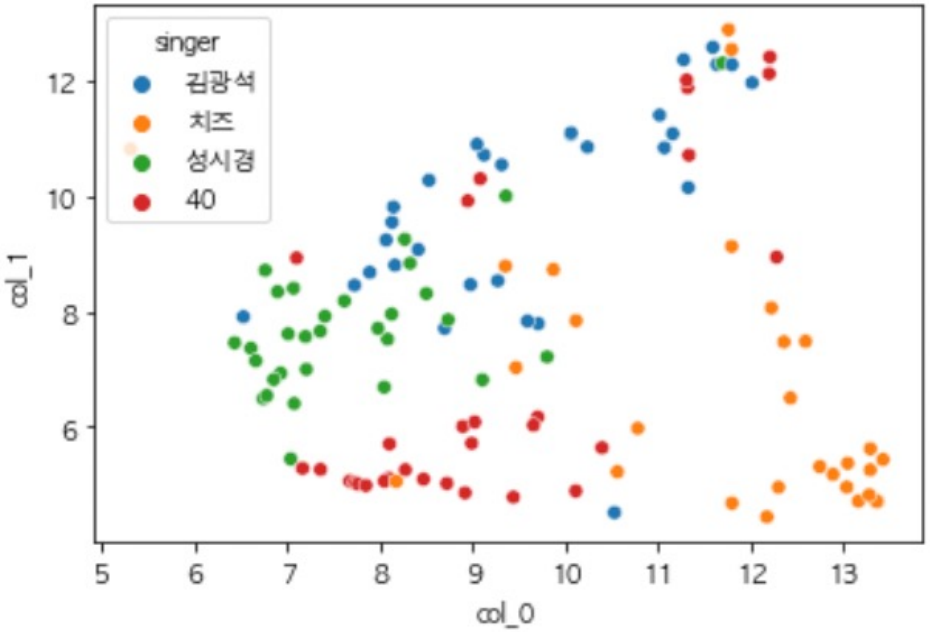
실험 결과 : Contrastive learning

Pitch Shift : neg

Time stretch : neg



성별 별 분포 시각화



장르가 다른 가수 4명의 분포 시각화

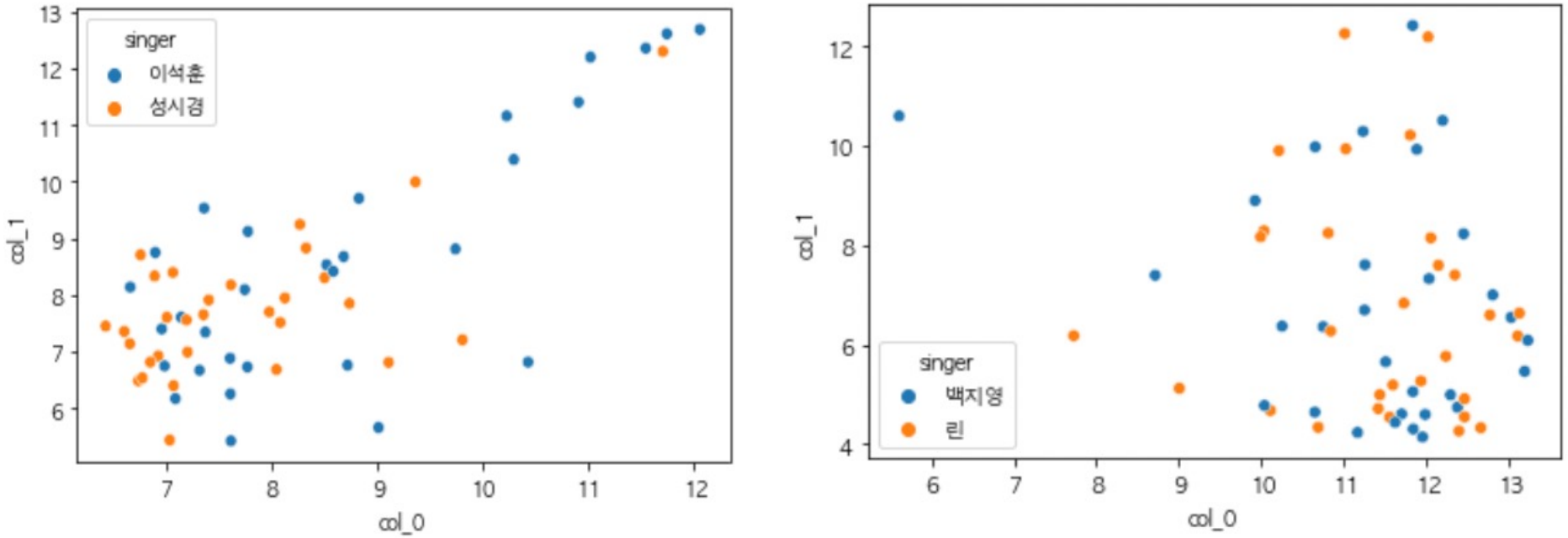
03

Experiments

실험 결과 : Contrastive learning

Pitch Shift : neg

Time stretch : neg



성별과 장르가 같고 음색과 보컬 스타일이 유사하다고 판단되는 가수들이 비슷한 분포를 나타냄

→ **Extract된 Feature representation이 음색과 보컬 스타일을 잘 반영**하고 있음을 보여줌

04

진행계획

	상황	진행 완료	진행 계획
Gender Classification	△	구현 중	~ 11/13 실험
Contrastive Learning	○	Experiment, Extract 완료	
TCAE	△	구현 완료 및 실험	성능 개선을 위한 구조 변경
RecSys	△		~ 11/13 구현

Reference

Liu, Jinglin, et al. "Learning the Beauty in Songs: Neural Singing Voice Beautifier." arXiv preprint arXiv:2202.13277 (2022).

Yakura, Hiromu, Kento Watanabe, and Masataka Goto. "Self-Supervised Contrastive Learning for Singing Voices." IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2022): 1614-1623.

Oord, Aaron van den, et al. "Wavenet: A generative model for raw audio." arXiv preprint arXiv:1609.03499 (2016).

감사합니다.

이지평 장성현 김보현 김종윤 김정하