



# 너의 다음곡이 보여

: Self Supervised Learning을 이용한  
음색 기반 노래 추천 시스템

이지평 장성현 김보현 김종윤 김정하



# 팀원 소개



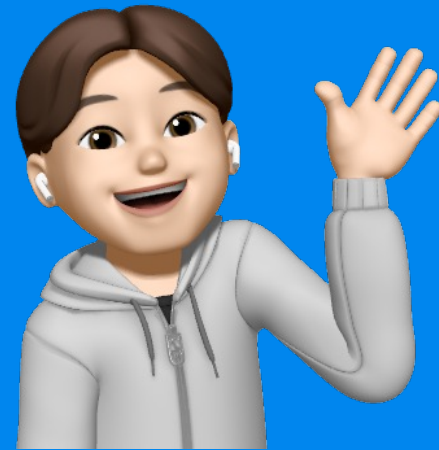
이지평



장성현



김보현



김종윤



김정하



# INDEX



## 주제 선정 배경

- 주제 선정 배경
- Framework



## Contrastive Learning

- Self-supervised learning
- Self supervised Contrastive learning for singing voices
- MoCo
- NCE loss
- Experiment



## TJAE

- Time Jump Auto Encoder
- Dilated Causal Convolution
- Experiment



## RecSys

- 콘텐츠 기반 추천 시스템
- Improvement & Future work

01



01

## 주제 선정 배경



다음에는 뭘 불러야 하지?

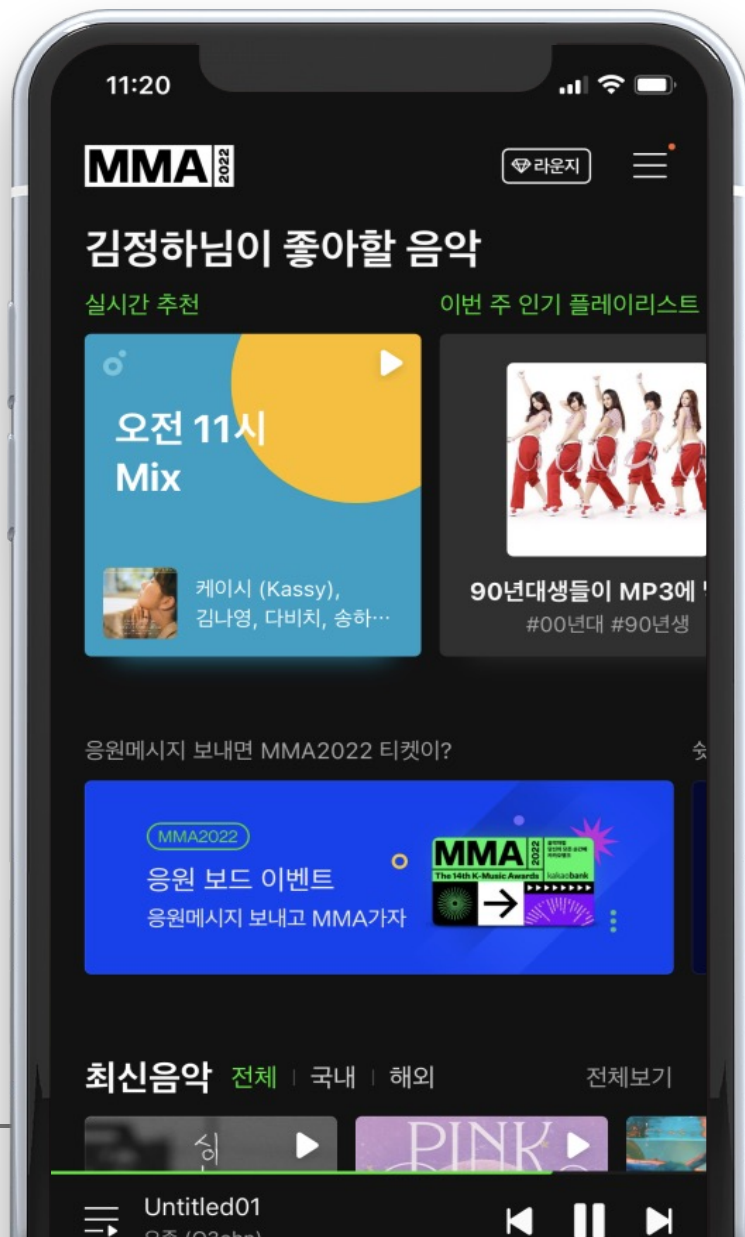
나에게 잘 어울리는 곡은?

부르는 노래 추천 시스템은 없을까?

The screenshot shows the TJ Music website interface. At the top, there's a navigation bar with links like 'TJ 노래방', '제품소개', '고객서비스', 'PR', 'IR', '회사소개', and '채용정보'. Below this, the main content area is titled '노래방 인기곡' (Karaoke Popular Songs). It features a search bar with filters for '가요' (Pop), '2022년 11월 01일 ~ 2022년 11월 20일', and a '검색하기' (Search) button. A message indicates that the data is being updated and the search time may be longer. Below the search bar, a table lists popular songs with columns for '순위' (Rank), '곡번호' (Song Number), '곡제목' (Song Title), and '가수' (Artist).

순위	곡번호	곡제목	가수
1	81425	사건의지평선	윤하
2	81405	내가아니라도	주호
3	12291	Monologue	버즈
4	78872	첫사랑	백아
5	82308	Monologue	tei
6	82466	Nxde	(여자)아이들

On the left side of the page, there's a sidebar with a dropdown menu for 'TJ 반주곡' (TJ Accompanying Songs) and several links: '노래검색', '이달의 신곡', '노래방인기곡', '노래신청', 'TJ 이벤트', 'TJ 반주기 활용팁', '아프리카TV 라이브 방송', '터치노래방', and '클럽노래방'.



주제 선정 배경

## 기존 추천시스템의 한계

Music

RecSys

음악 추천시스템에 대한 기존 연구들은  
사용자의 인구통계학 정보,  
개인취향 및 최근 관심 곡 등을 기반으로  
듣는 음악에 대한 추천이 대부분

Point.1

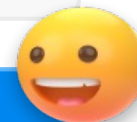
듣는 노래가 아닌 부르는 노래를 추천할 수 있을까?

Point.2

음색을 기반으로 부르는 노래를 추천하면 어떨까?

Point.3

사용자의 목소리만으로 추천을 해주는 시스템을 만들자!



01

## 주제 선정 배경

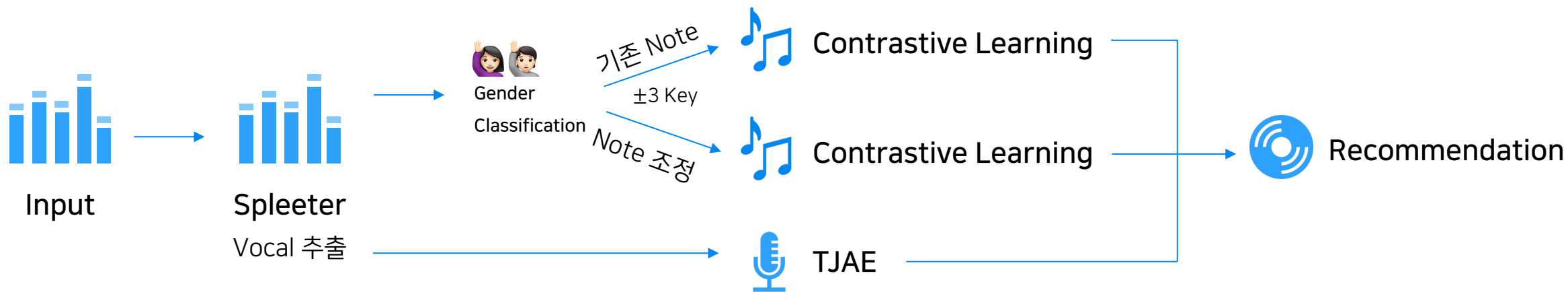


“ 듣는 음악 추천이 아닌 부르는 음악 추천 ”

사용자의 목소리 하나만으로 어울리는 노래를 추천 해주는  
콘텐츠 기반의 추천시스템 모델을 개발하고자 함

## 01

## Framework





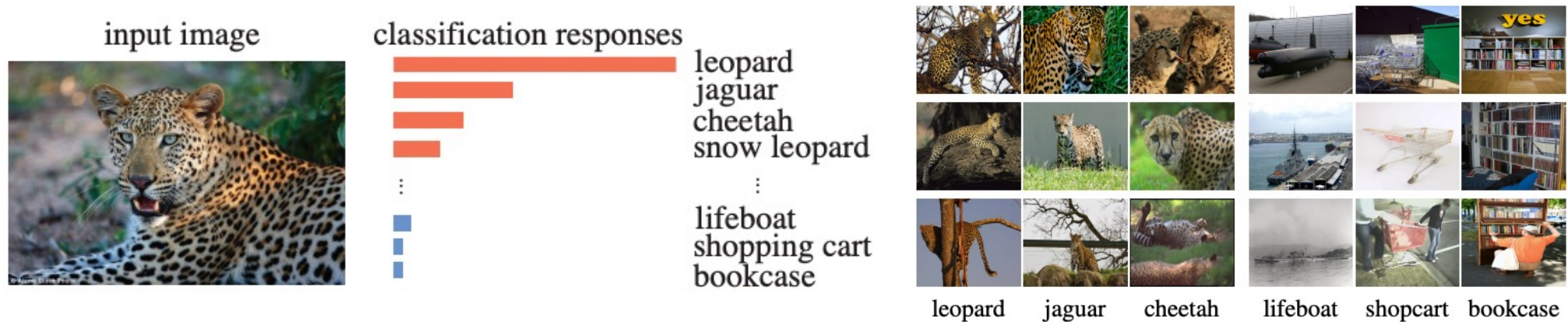
## 02

# Contrastive learning

Self-supervised learning

## Self-supervised learning의 기본 아이디어

“레이블 정보 없이 관측치 레벨에서 학습 및 공유하는 representation, semantic structure가 있을 것이다.”

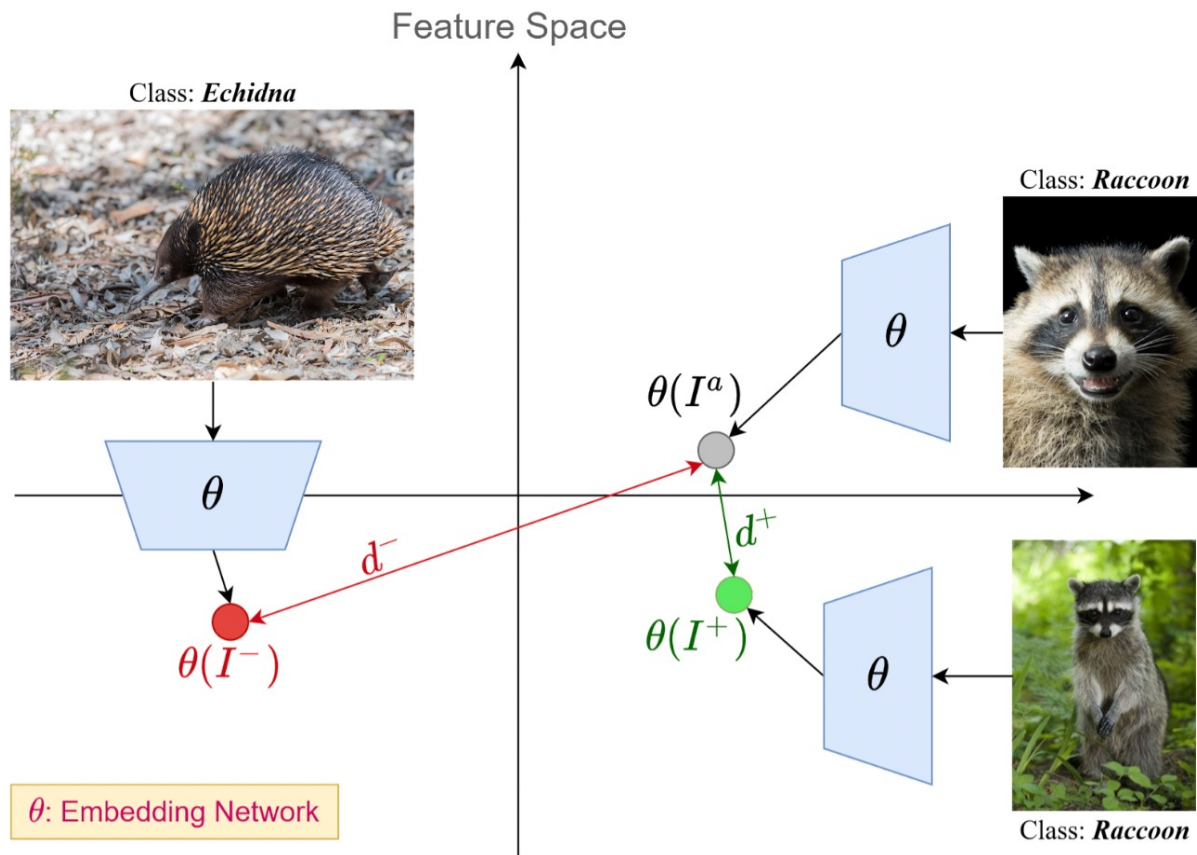


- Supervised learning에서 semantic labeling이 아님에도 불구하고 확률값이 높은 클래스는 타겟 클래스와 시각적으로 유사
- 관측치 사이의 유사성을 기반으로, 관측치끼리 구분하도록 학습을 한다면 레이블 정보 없이도 좋은 representation을 얻을 수 있음

## 02

# Contrastive learning

Self-supervised learning



## Contrastive learning

Contrastive learning의 기본 Frameworks는

anchor라고 하는 데이터 샘플,

"positive" sample이라고 하는,

anchor와 동일한 분포에 속하는 데이터 포인트와

"negative" sample이라고 하는

다른 분포에 속하는 또 다른 데이터 포인트의 선택으로 구성

잠재 공간에서 anchor와 positive sample 간의 거리를 최소화하고

동시에 anchor와 negative sample 간의 거리를 최대화

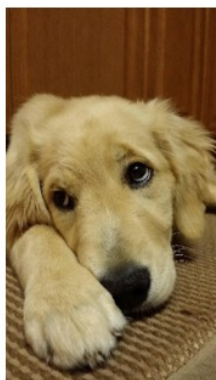
## 02

# Contrastive learning

Self-supervised learning

## 기존 Contrastive learning

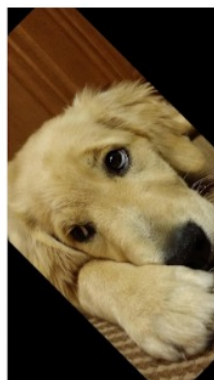
다양한 방식으로 데이터를 Augmentation해서 **positive pair**를 만들어 냄



Original



Color Jitter



Rotation



Flipping



Noising



Affine

## 02

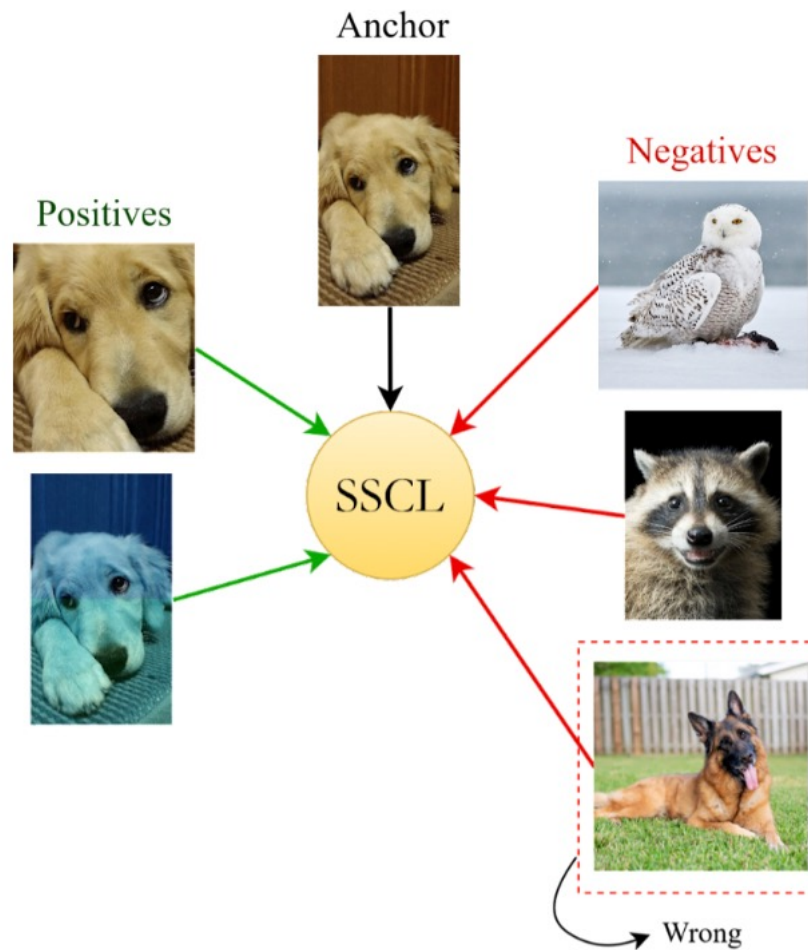
# Contrastive learning

Self-supervised learning

기존 Contrastive learning

**Positive pair** : Anchor의 augmentation한 결과

**Negative pair** : 다른 데이터 샘플

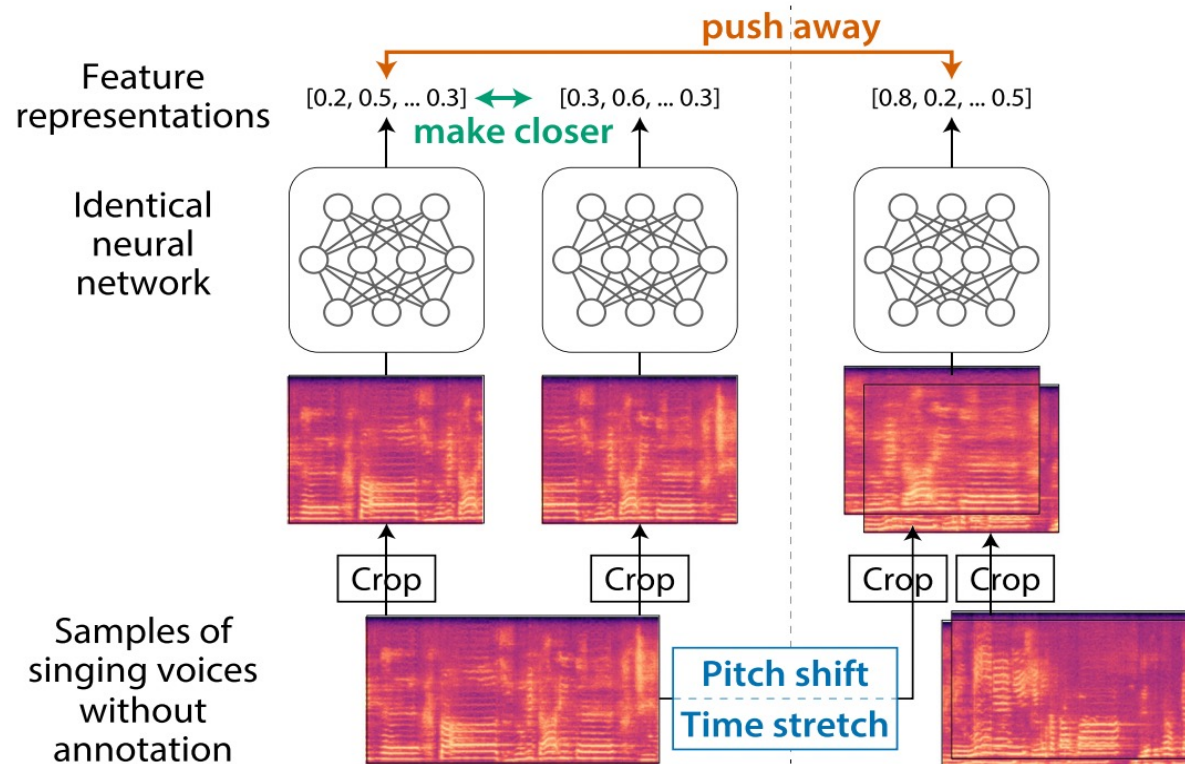


## 02

## Contrastive learning

## Self-supervised Contrastive Learning for Singing Voices

Singing Voices에서 Vocal timbre, Singing expression을 반영한 Feature representation



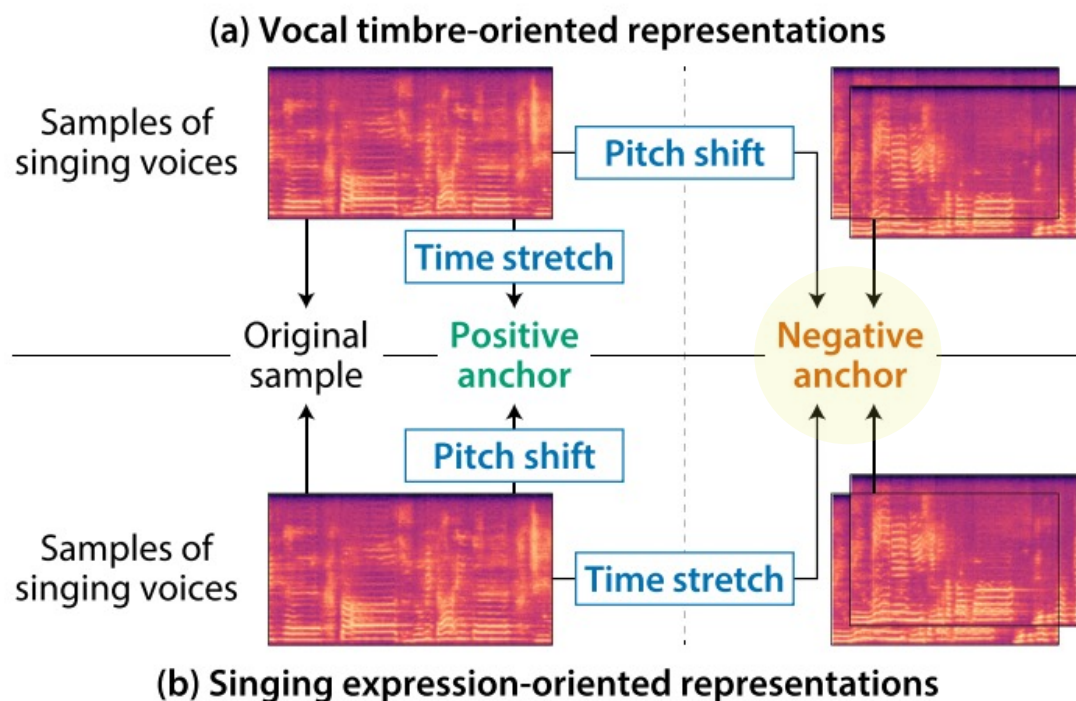
(b) Proposed approach



## 02

# Contrastive learning

Self supervised Contrastive learning for singing voices



한 Anchor에서 Augmentation한 결과를  
얻고자 하는 representation에 따라  
positive pair 혹은 negative pair로 설정

Vocal timbre를 반영한 Feature representation을 얻고자 할 때,

Pitch shift → Negative Pair

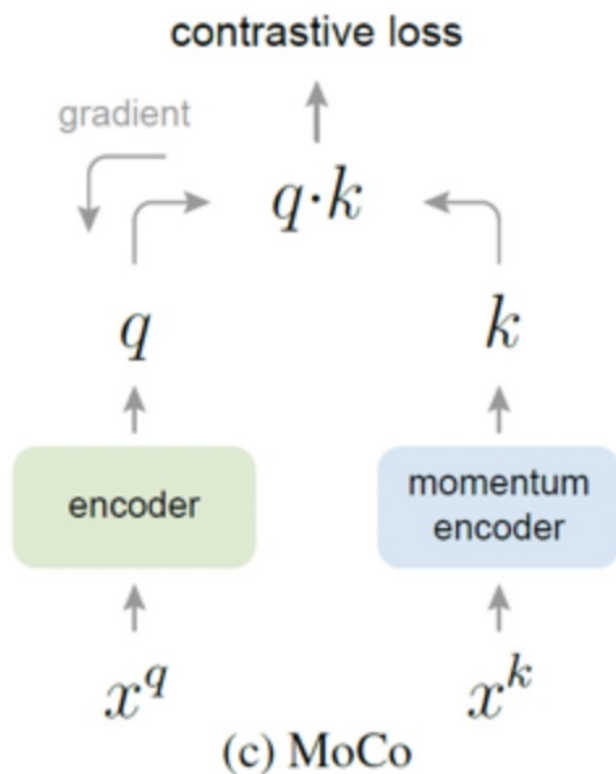
Singing expression을 반영한 Feature representation을 얻고자 할 때,

Time stretch → Negative Pair

## 02

## Contrastive learning

MoCo



## Momentum Encoder

$$\theta_k \leftarrow m\theta_k + (1 - m)\theta_q$$

Query encoder와 Key encoder가 별도로 정의

Query encoder  $\theta_q$ , Key encoder  $\theta_k$  를 동일하게 설정하고 momentum update를 적용하여

Key encoder  $\theta_k$  를 천천히 업데이트

Key encoder는 query encoder에 대해 점진적으로 변하게 되어

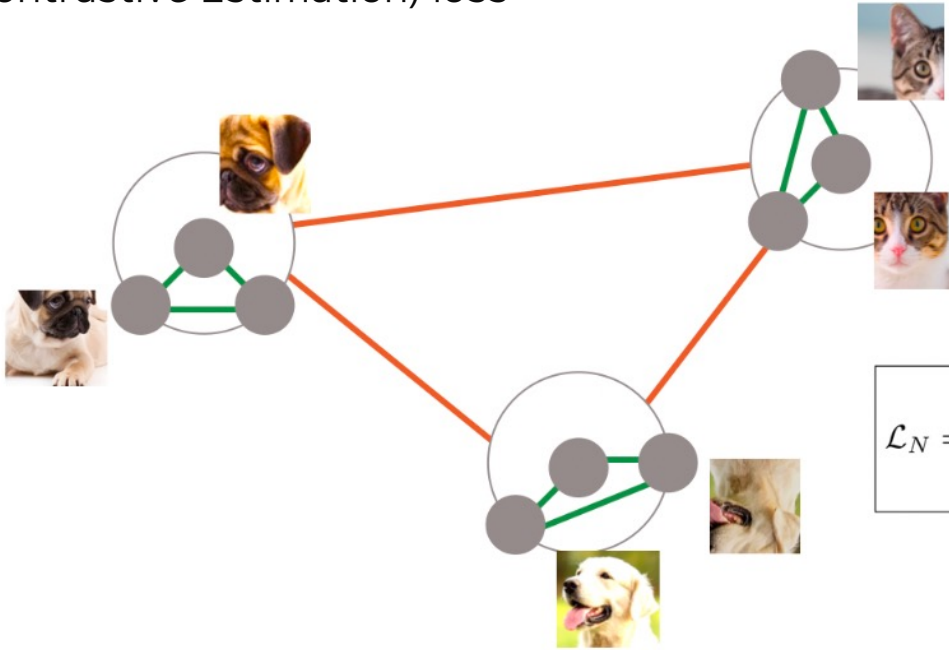
최대한 일관된 representation을 dictionary Key에 담을 수 있게 하고

Queue를 사용하여 dictionary를 최대한 크게 구성해서 많고 다양한 negative sample을 볼 수 있도록 함

## 02

# Contrastive learning

NCE(Noise Contrastive Estimation) loss



$$\mathcal{L}_N = -\mathbb{E}_X \left[ \log \frac{\exp(f(x)^T f(x^+))}{\exp(f(x)^T f(x^+)) + \sum_{j=1}^{N-1} \exp(f(x)^T f(x_j))} \right]$$

$$\text{score}(f(x), f(x^+)) \gg \text{score}(f(x), f(x^-))$$

Positive samples

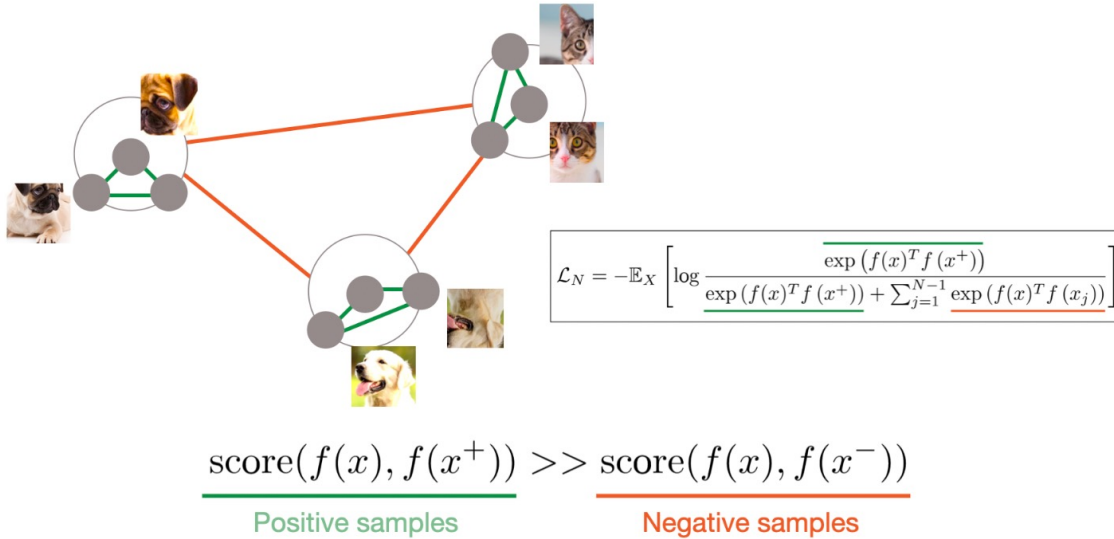
Negative samples



## 02

# Contrastive learning

NCE(Noise Contrastive Estimation) loss



## InfoNCE Loss

초록색은 positive sample과의 관계,  
빨간색은 negative sample과의 관계를 의미

InfoNCE Loss를 minimize하면

분자는 maximize되고, 분모는 minimize

score는 보통 cosine similarity 사용( InfoNCE Loss에서도 마찬가지 )

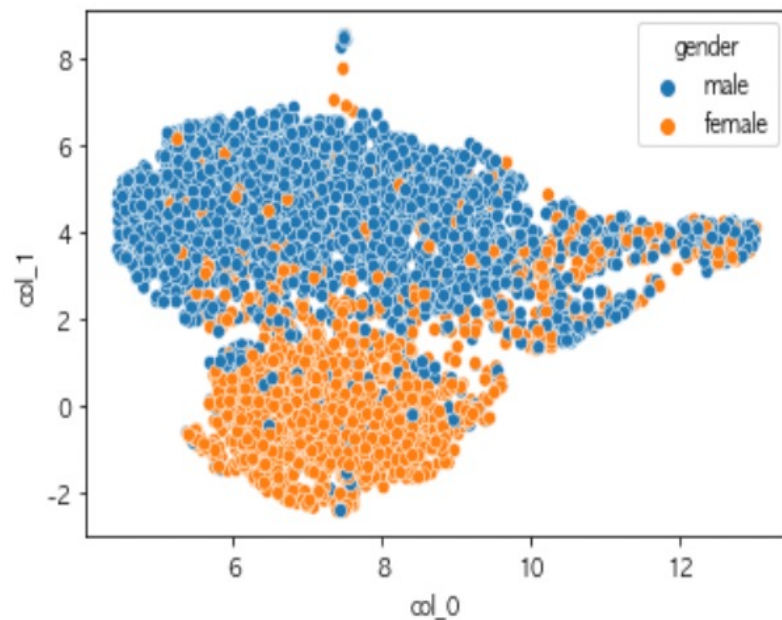
augmentation을 한 같은 이미지 간은 **Positive sample**로 거리가 가깝고,  
다른 이미지 간은 **Negative sample**로 거리가 먼 형태의 representation 형성

## 02

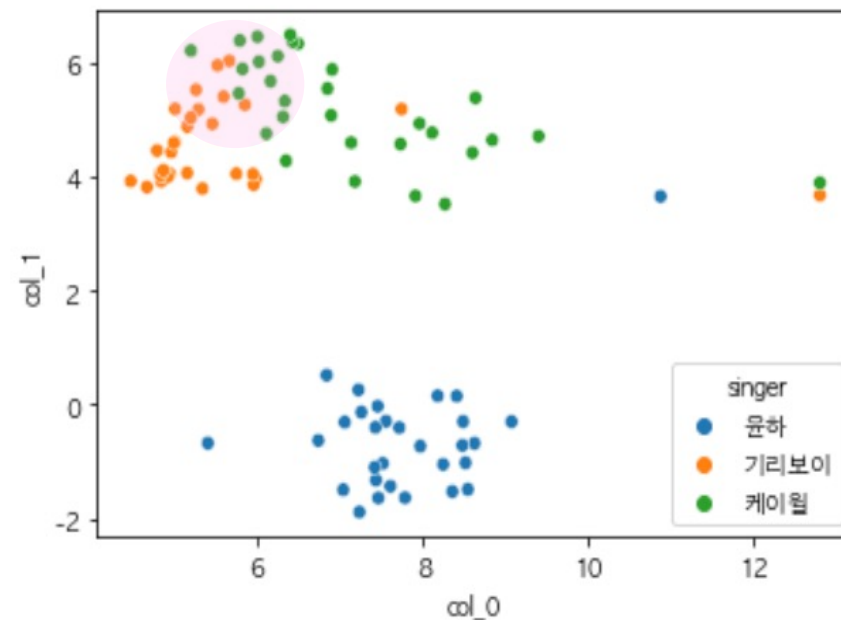
## Contrastive learning

Experiment

Pitch Shift : neg  
Time stretch : pos



성별에 따른 음색 차이 확인



가수와 장르에 따른 음색 차이 확인

→ Extract된 Feature representation이 해당 특징을 잘 반영하고 있음을 보여줌

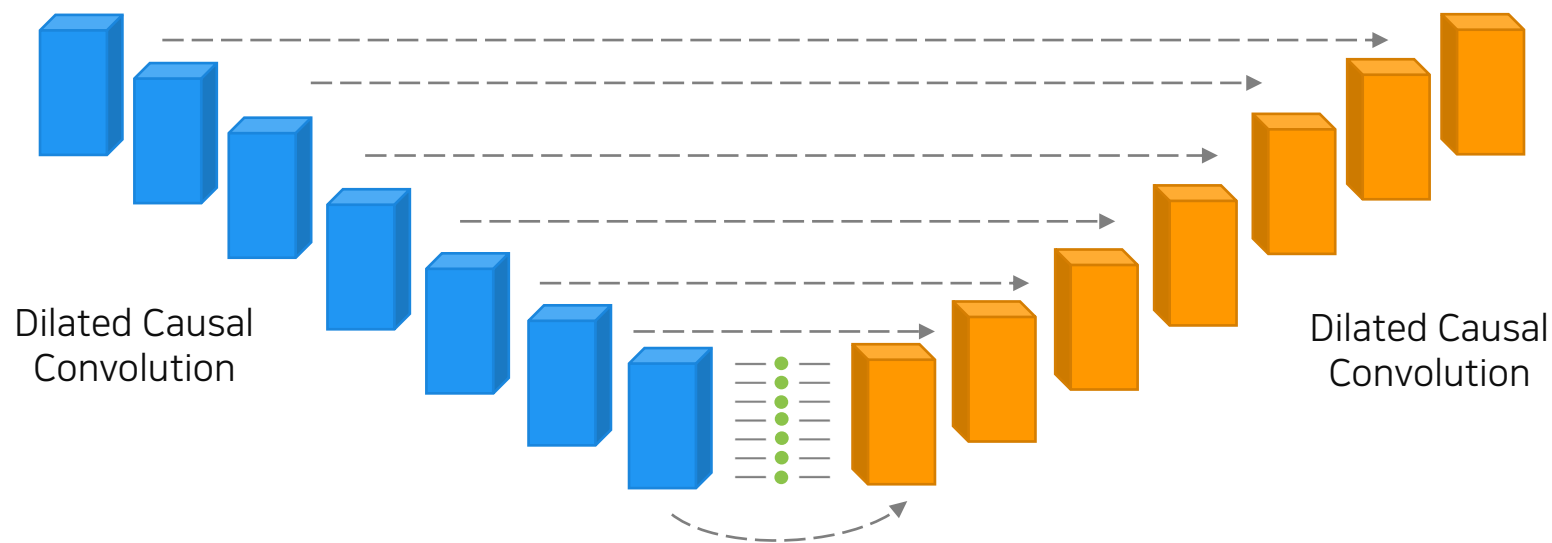
## 03

TJAE<sub>(KYAE)</sub>

Time Jump Auto Encoder

## TJAE

Contrastive learning을 통해 반영할 수 있는 Timbre, vocal style과 더불어,  
Vocal의 특징을 더욱 풍부하게 활용하고자 전체적인 vocal의 분위기를 반영할 수 있는 TJAE 구조 개발



## 03

## TJAE

## Dilated Causal Convolution

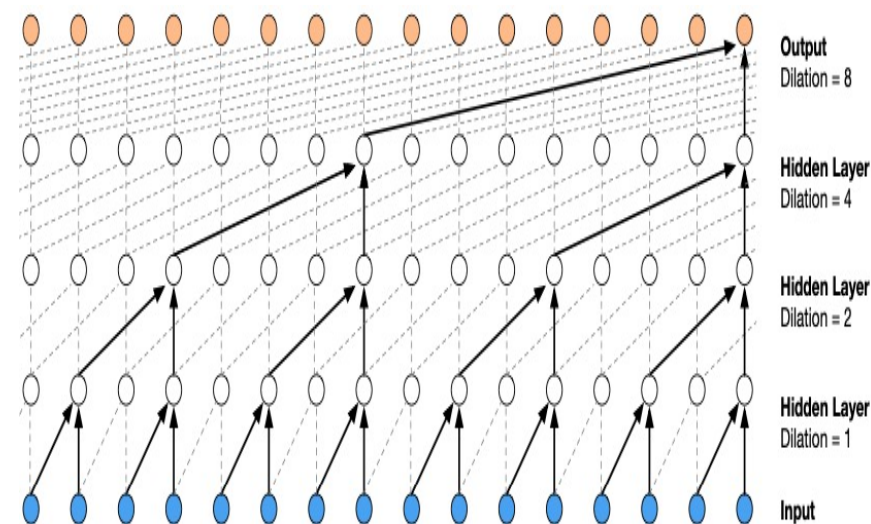
Dilated Causal Convolution = Causal Convolution + Dilated Convolution

### Causal Convolution :

시간 순서를 고려하여 Convolution Filter를 적용하는 변형 Convolution Layer  
Causal Convolution을 위로 쌓을수록 Input 데이터의 Receptive Field가 커짐  
→ 음성 데이터 모델링 가능

### Dilated Convolution :

추출 간격(Dilation)을 조절하여 더 넓은 receptive field를 갖게 하는 Convolution Layer  
Receptive field를 넓히기 위해 많은 양의 Layer를 쌓아야 하는  
causal convolution의 단점을 극복할 수 있음  
→ 상대적으로 적은 양의 layer로 receptive field를 넓히는 효과



## 03

## TJAE

Time Jump Auto Encoder

## TJAE

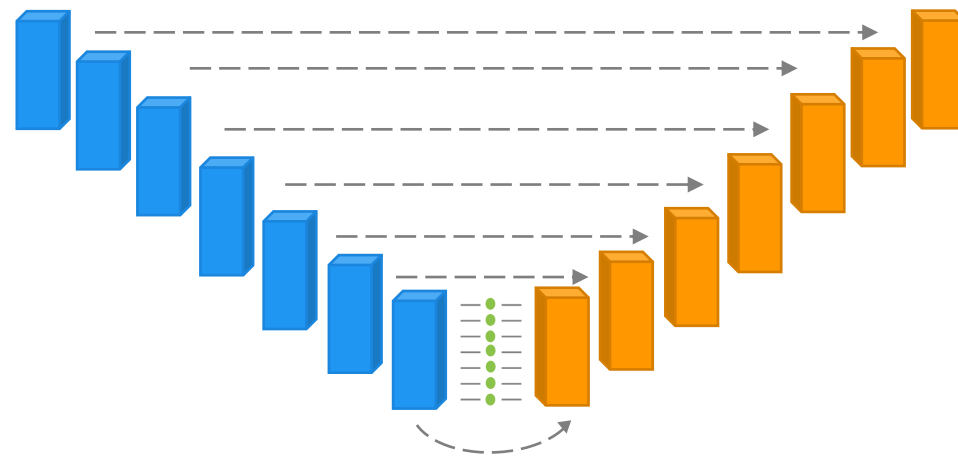
Autoencoder를 기반으로 Skip Architecture를 구축

Autoencoder의 각 층은 Dilated를 다르게 한 Convolution Layer로 구성

다른 시간 의존성을 가진 Feature map을 decoder의 층에 전달

Decoder가 층마다 다른 시간의 정보를 추가로 가짐으로써 Input을 재구축

이를 통해 생성된 Latent Space는 분위기의 정보를 가짐



## 03

## TJAE

## Experiment

## TJAE

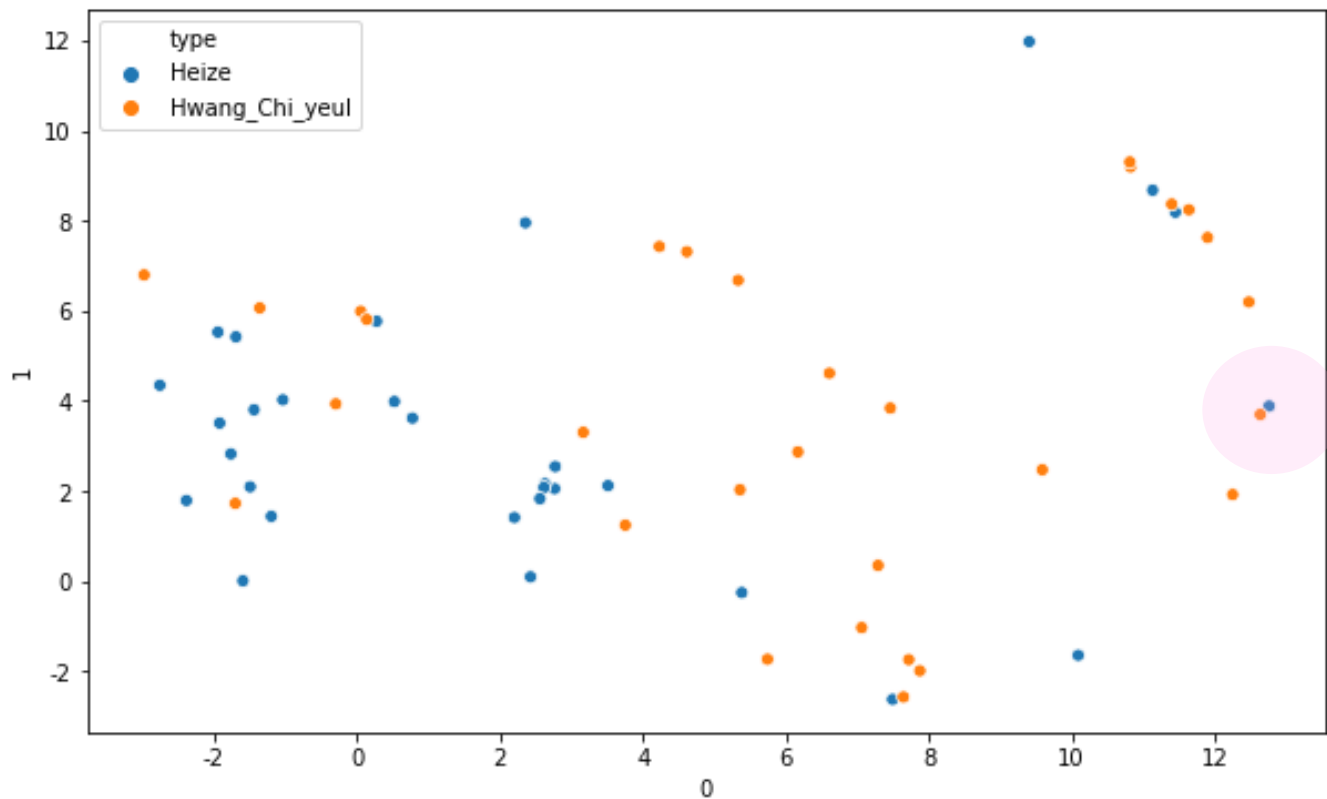
Autoencoder를 기반으로 Skip Architecture를 구축

Autoencoder의 각 층은 Dilated를 다르게 한 Convolution Layer로 구성

다른 시간 의존성을 가진 Feature map을 decoder의 층에 전달

Decoder가 층마다 다른 시간의 정보를 추가로 가짐으로써 Input을 재구축

이를 통해 생성된 Latent Space는 분위기의 정보를 가짐



04

## RecSys

콘텐츠 기반 추천 시스템

### - 앞서 생성한 여러 가지 Embedding

#### Embedding by Self Supervised Contrastive Learning

Pitch-shifted (neg) + Time-stretched (pos)

: Vocal Timbre에 집중해서 학습

Pitch-shifted (neg) + Time-stretched (neg)

: Vocal Timbre & Singing Expression 모두에 집중해서 학습

#### Embedding by TJAE ( Time Jump Auto Encoder )

전체적인 분위기, 느낌에 집중해서 학습

( 확실히 위에 두 개의 Embedding과는 다름 )

## 04

## RecSys

콘텐츠 기반 추천 시스템



Note 조정

SSCL Emb

Pitch Shift : Neg

Time Stretch : Pos



Note 조정

SSCL Emb

Pitch Shift : Pos

Time Stretch : Neg

SSCL Emb

Pitch Shift : Neg

Time Stretch : Pos

SSCL Emb

Pitch Shift : Pos

Time Stretch : Neg

TJAE Emb

5가지 Embedding 각각의 코사인 유사도를 가중합하여 최종 유사도 생성  
가중치는 여러 실험 끝에 heuristic하게 정함



04

## RecSys

콘텐츠 기반 추천 시스템



Input



Recommendation

## 04

## Improvement & Future work

### ♥ Improvement ♥

1. 선행 연구 부족
2. 학습 데이터 셋 안정성 및 절대적 수량 부족
3. 일정 길이 이상의 Input 필요(120s)
4. Mac과 Window 호환성 이슈가 생각보다 쉽지 않음
5. 컴퓨팅 파워 부족
6. 수면 부족
7. 예산 부족

### ♥ Future work ♥

1. 학습 데이터 셋 안정성 확보 및 추가 데이터 수집
2. 단순 Embedding 유사도가 아닌 추천시스템 모델 활용
3. 성별 예측 단계 추가
4. 구체적 특징 추출가능한 서브 모델링 추가
5. 설문조사를 이용한 통계적 정량 지표 생성

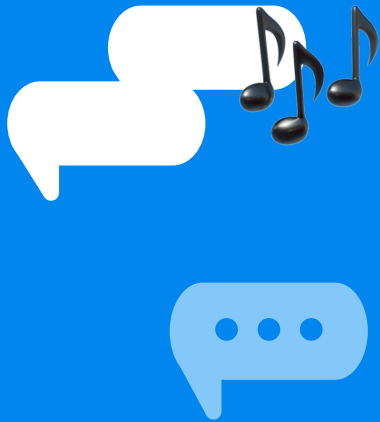


# Reference

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

