

# Advanced StarGAN-VC: Extracting precise speaker feature using speaker encoder from AutoVC

류동원

2018015350 류형진

2018015523 윤명현

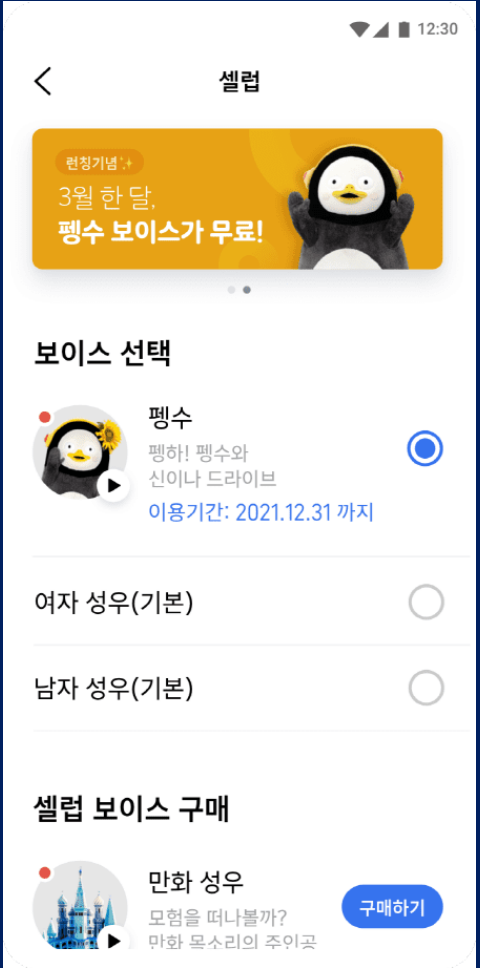
2020099689 전민주

2017 정휘준

2018015750 최승훈

# Introduction

<티맵 모빌리티 "티맵 셀럽" 서비스>



티맵모빌리티가 '티맵 셀럽' 서비스를 시작한다고 4일 밝혔다. 유명인이나 캐릭터의 목소리로 길 안내를 하는 식이다.

티맵 셀럽은 AI 기반 음성합성시스템(TTS) 기술을 활용한 서비스다. 셀럽의 유행어나 억양과 TTS 합성 음성을 혼합해 길안내를 제공한다.

EBS 인기 캐릭터 '펭수' 목소리를 시작으로 이달 중 안영미, 다음 달엔 애니메이션 겨울왕국의 안나 목소리를 연기한 성우 박지윤 등 연내 20여개 목소리가 추가될 예정이다.

2020년 12월 9일부터 엠넷에서 방영하는 AI 프로젝트 방송이다. 2부작으로 고인이 된 터틀맨과 김현식의 목소리와 모습 등을 최신 기술로 복원해 새로운 노래를 부르게 하는 주제이다. 진행자는 하하. 터틀맨 편이 공개되자 유튜브와 각종 커뮤니티 등에 빠르게 퍼져나갔다.

<AI 음악 프로젝트 다시 한 번,>



최근 다양한 방송 및 음악 업계, 다양한 매체에서  
**Voice Conversion** 기술을 활용한 서비스가 출시

# Introduction

## 음성 변환(Voice Conversion)

- **음성 변환(Voice Conversion)** 이란 Source Speaker의 음성을 Target Speaker의 음성으로 변환하는 것.
  - ✓ ① 음성에서 언어적 내용(linguistic contents)은 변하지 않고,
  - ✓ ② 화자의 음성 특징(리듬, 음역, 음색, ...)만을 변환하는 것.

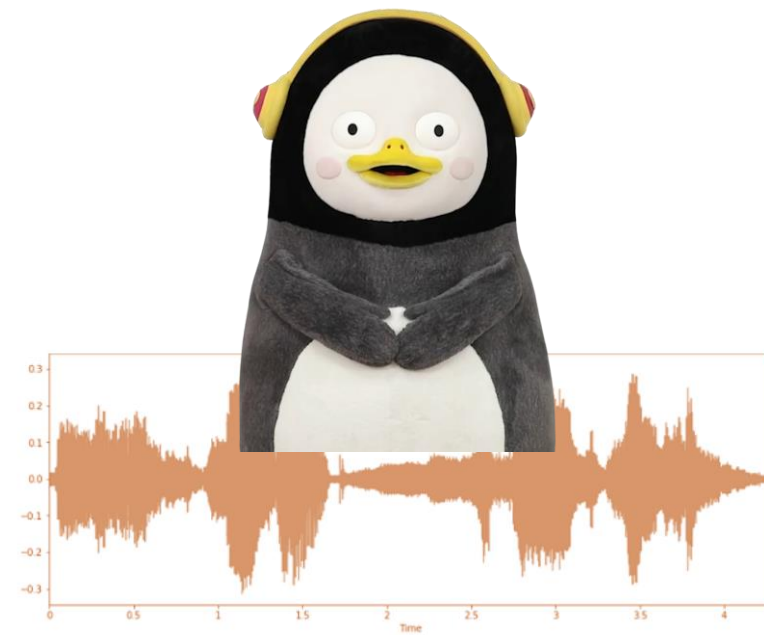


**Source Speaker**  
 "(내 목소리로) 한양대로 가자"



**Voice Conversion**

언어적 내용 변화 X  
 화자의 음성 특성만을 변환

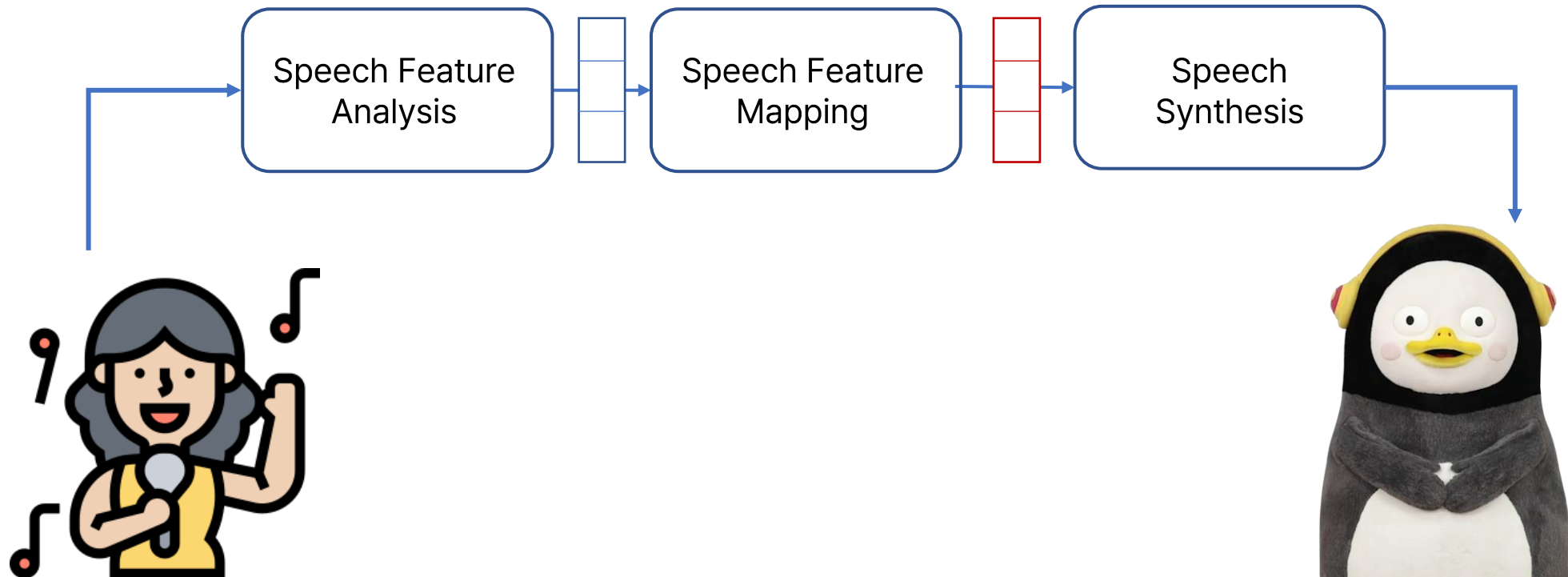


**Target Speaker**  
 "(펭수 목소리로) 한양대로 가자"

# Introduction

## Voice Conversion 진행과정

- ① 음성에서 화자의 특징 및 발화 내용이 포함되어 있는 Feature를 추출
- ② Source 화자의 특징 Feature를 Target 화자의 특징 Feature로 변환
- ③ 변경된 Feature를 활용하여 Target 화자의 음성으로 합성

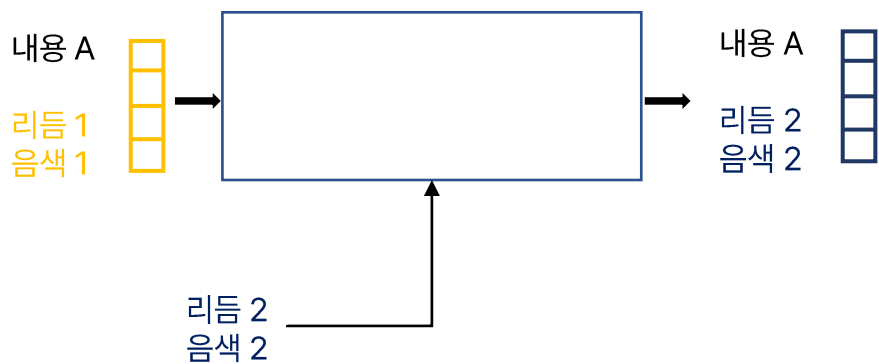


Voice  
Conversion

## Previous Research

- 모델의 구조에 따라 : Direct Conversion vs Feature Disentangle

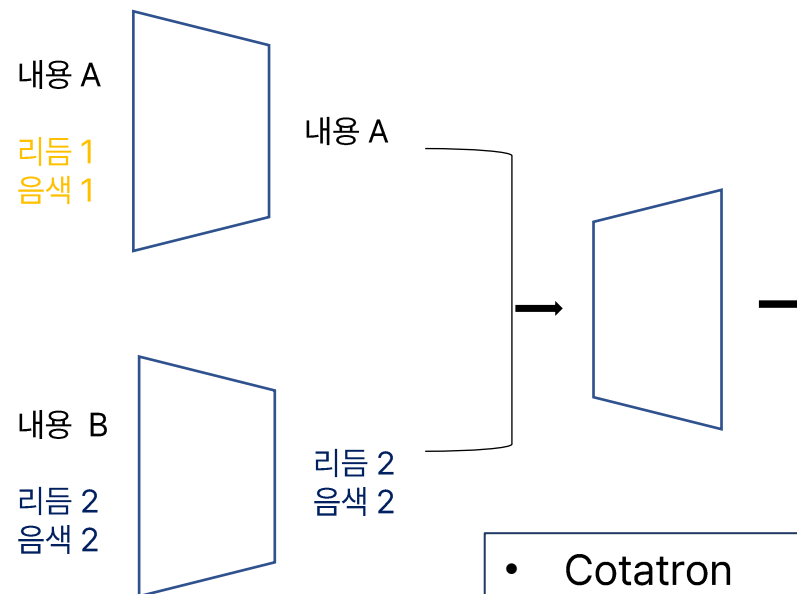
## ① Direct Conversion



- CycleGAN-VC
- StarGAN-VC

→ 빠른 속도, 적은 데이터로 높은 품질 추출

## ② Feature Disentangle



- Cotatron
- Fragment-VC
- AutoVC

→ 간단한 구조, zero-shot 가능

## Previous Research

- 모델의 구조에 따라 : Direct Conversion vs Feature Disentangle

## ① Direct Conversion

내용 A

리듬 1  
음색 1리듬 2  
음색 2

Direct Conversion 방식은  
데이터 처리 과정에서의 시간이 덜 소모되므로  
분석의 성능이 높고,  
Feature Disentangle 방식은  
Many to Many 화자와 Zero shot conversion이 가능 하므로  
넓은 활용범위를 가지는게 특징

두가지의 장점을 모두 가지게 된다면 성능이 더 향상될 것으로 판단

- CycleGAN-VC
- StarGAN-VC

→ 빠른 속도, 적은 데이터로 높은 품질 추출

## ② Feature Disentangle

내용 A

리듬 2  
음색 2

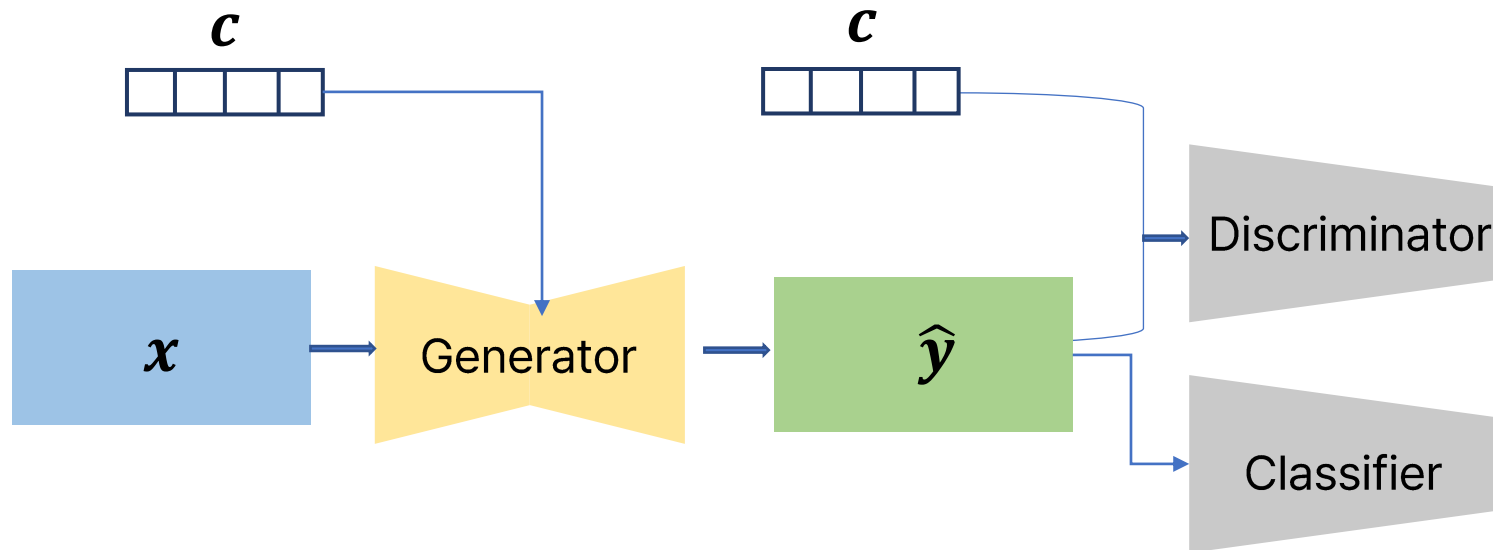
- Cotatron
- Fragment-VC
- AutoVC

→ 간단한 구조, zero-shot 가능

## Base Model

### StarGAN-VC

- Non-parallel many-to-many voice conversion with star generative adversarial networks, 2018(IEEE)
- GAN 계열의 아키텍처를 제안한 논문으로 CycleGAN-VC 의 단점을 극복하기 위한 장치를 포함.
  - 여러명의 사람 음성을 합성할 수 있으며, World Vocoder와 함께 활용하여 실시간(빠른) 음성합성이 가능.
  - 비교적 적은 Non-parallel 데이터를 이용하여 실제 음성과 비슷한 음성을 생성할 수 있음



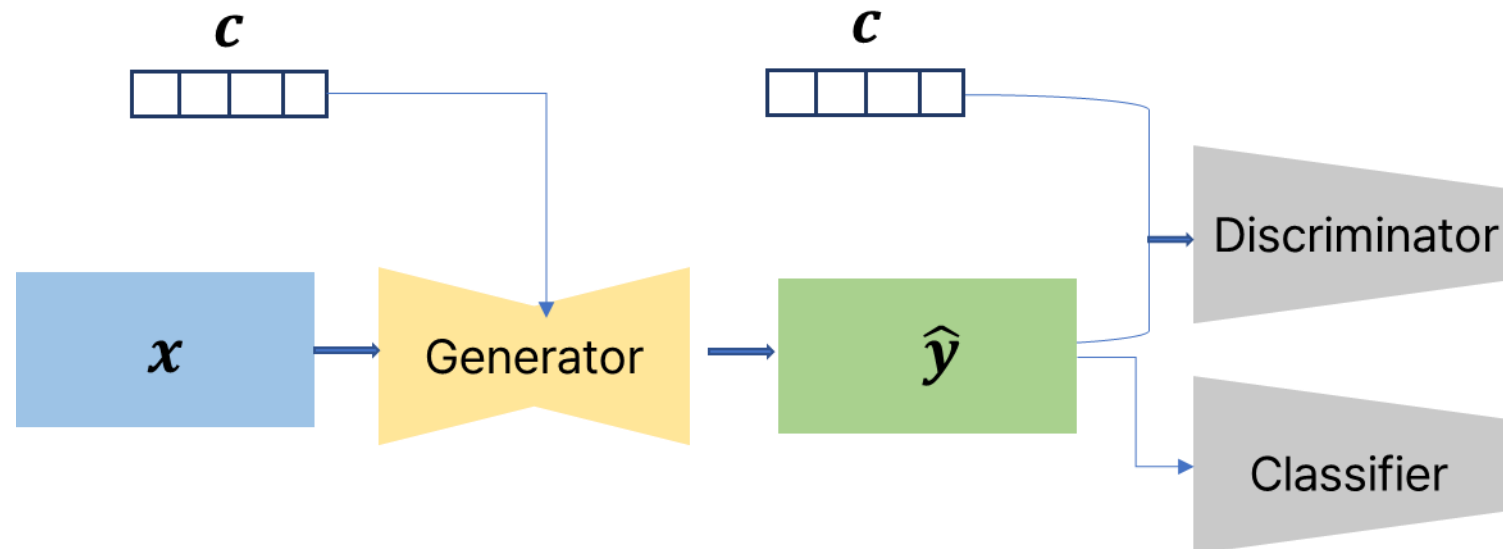
## Base Model

### StarGAN-VC

- Non-parallel many-to-many voice conversion with star generative adversarial networks, 2018(IEEE)

① Discriminator | 실제 음성 높은 확률, Fake 음성 낮은 확률 도출되도록 학습

② Classifier | 음성 Feature 통해 해당 음성이 누구의 음성인지 구분할 수 있도록 학습



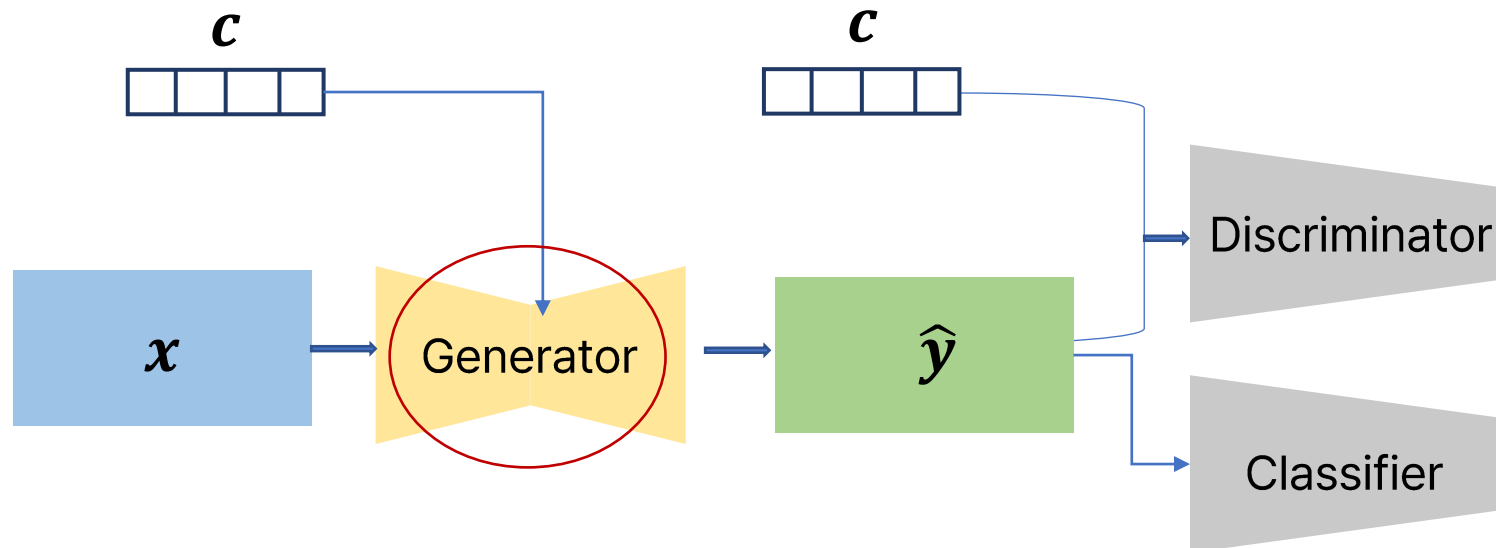


## Base Model

### StarGAN-VC

- Non-parallel many-to-many voice conversion with star generative adversarial networks, 2018(IEEE)

- ③ Generator | 1) Generator에서 생성된 음성 Feature가 Real/Fake Discriminator 속일 수 있도록 학습  
2) Generator에서 생성된 음성 Feature 가 Target 음성특징 가질 수 있도록 학습  
3) Generator 에서 생성된 음성 Feature 의 언어정보가 보존될 수 있도록 Cycle Consistency 학습  
4) Target(c) 와 Source(c)가 동일할 때 Generator 가 동일한 음성 생성할 수 있도록 학습



## Base Model

### StarGAN-VC

- Non-parallel many-to-many voice conversion with star generative adversarial networks, 2018(IEEE)

#### ✓ 장점

- ① Parallel 데이터 (여러명의 화자가 동일한 내용을 이야기한 데이터)가 없어도 학습이 가능한 아키텍처
- ② 아키텍처를 CNN 계열의 구조의 모델로 구성했기 때문에 실시간 음성변환이 가능

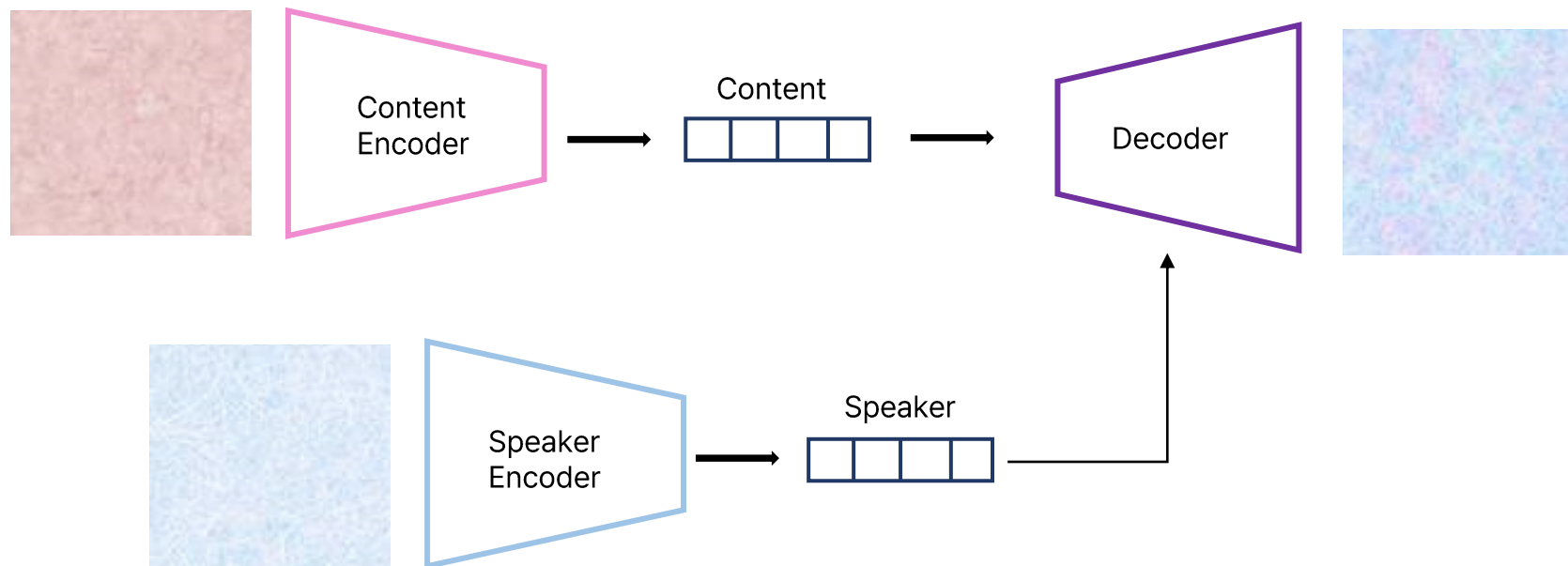
#### ✓ 단점

- ① 파라미터의 영향을 많이 받기 때문에 좋은 음성을 생성하기 위해서 많은 튜닝이 필요함

## Base Model

### Auto-VC

- Zero-Shot Voice Style Transfer with Only Autoencoder Loss, 2019(ICML)
1. 간단한 Auto-Encoder 구조를 활용하여 음성의 Content와 Local Feature를 분리.
  2. Self-Reconstruction Loss만을 활용하여 모델을 학습하므로 안정적임.
  3. 순수한 품질의 Zero-shot Conversion(몇 개의 음성데이터만으로 음성변환)이 가능한 첫번째 모델.

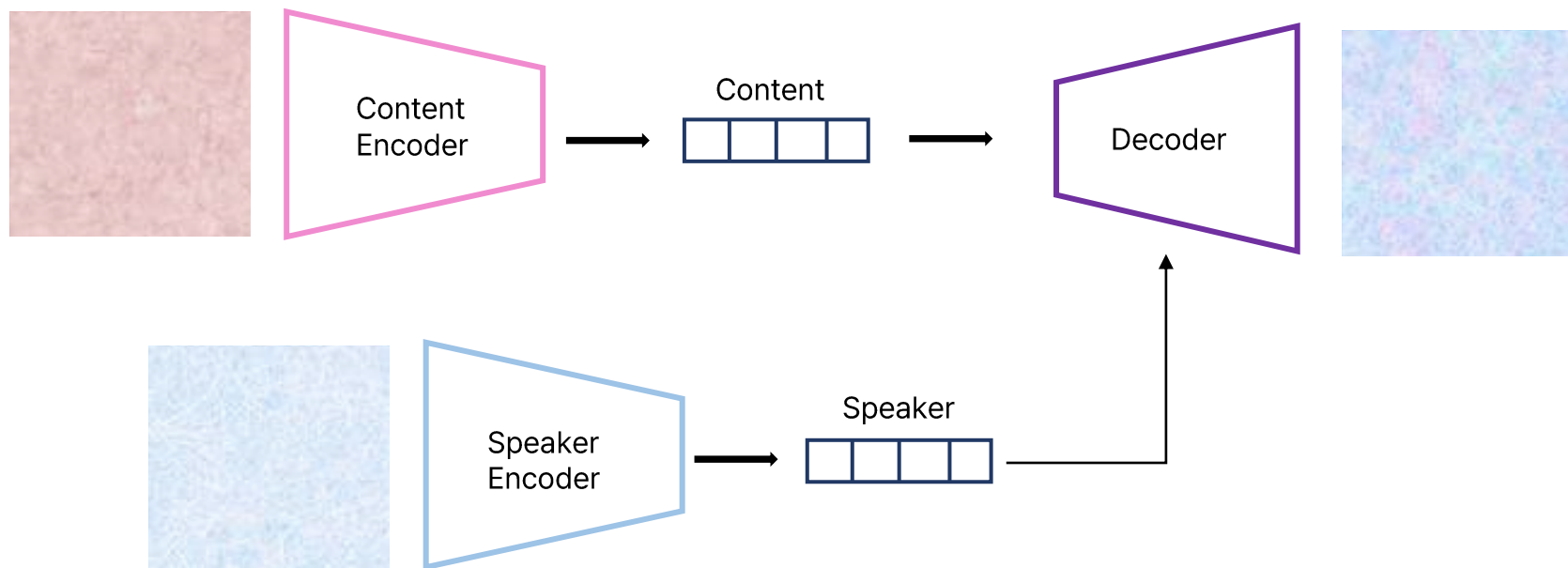


## Base Model

### Auto-VC

#### Architecture

1. Content Encoder : 음성 Feature에서 Content 벡터를 추출하는 모듈
2. Speaker Encoder : 음성 Feature에서 Speaker의 특징 벡터를 추출하는 모듈
3. Decoder : Content 벡터와 Speaker 특징 벡터를 입력으로 받아 Speaker의 특징이 포함된 음성 Feature를 생성하는 모듈



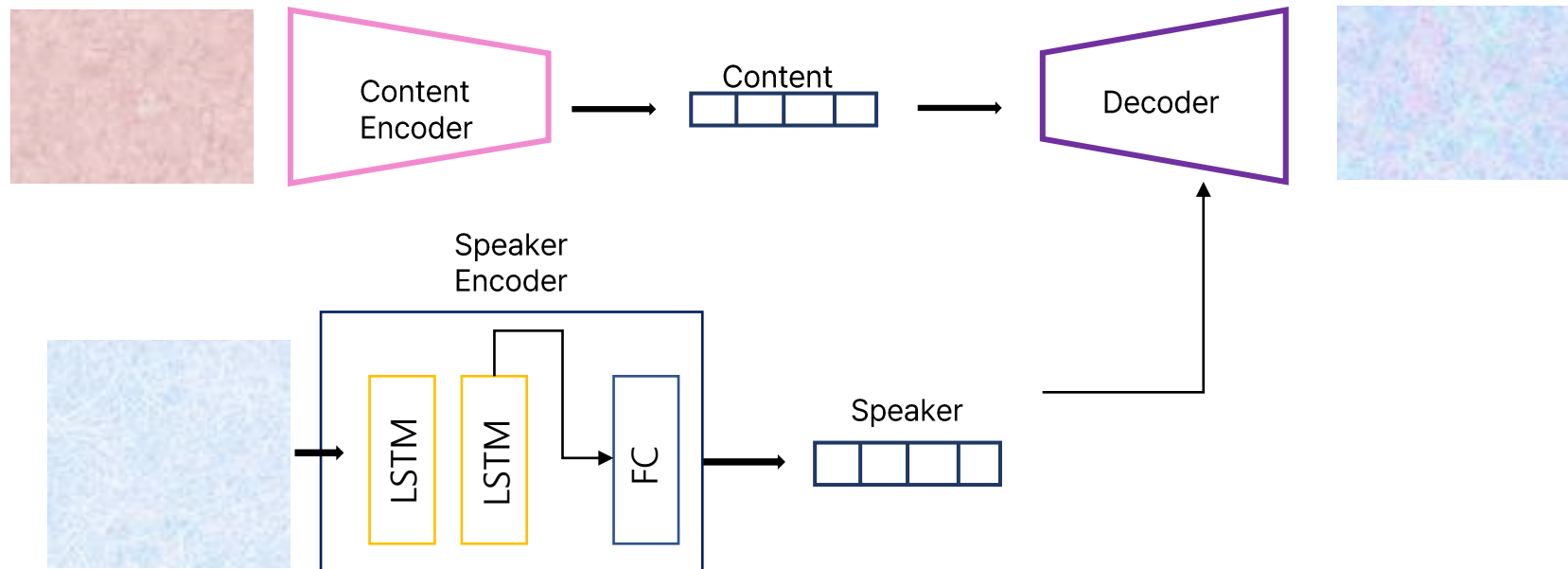
## Base Model

### Auto-VC

#### Speaker Encoder

길이가 가변적인 음성 Feature를 입력으로 받아 256 dim 크기의 Speaker feature를 출력해주는 모듈.

- ✓ (일관성) 동일한 화자의 음성을 Speaker Encoder에 넣었을 때 유사한 Speaker Feature를 추출할 수 있어야 함.
- ✓ (유사성) 비슷한 음색을 갖고 있는 두 화자의 Speaker Feature들은 서로 비슷해야 함.



## Base Model

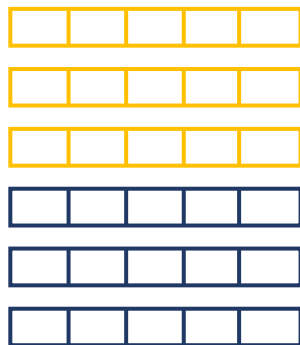
### Auto-VC

#### Speaker Encoder 학습방법 – Generalized End to End (GE2E)

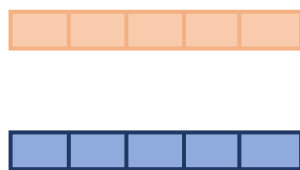
- ① 동일한 화자의 음성을 이용하여, 각각 Speaker Feature와 Centroid를 도출
- ② 각 Speaker Feature와 Centroid 거리를 cosine similarity를 활용하여 계산
- ③ 동일한 화자의 Centroid는 선택한 Speaker Feature가 가깝게, 다른 화자의 Centroid는 멀게 학습

$$L(e_{ji}) = -S_{ji,j} + \log \sum_{k=1}^N \exp(S_{ji,k})$$

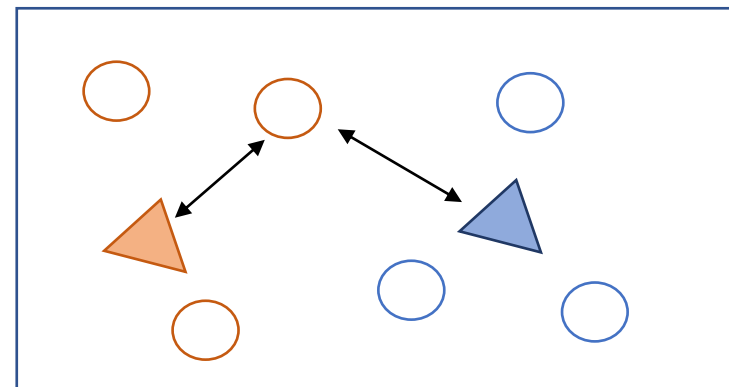
Speaker feature ( $e_{ji,k}$ )



Centroid( $c_k$ )



$$\text{거리} : S_{ji,k} = w \cdot \cos(e_{ji,k}, c_k) + b$$

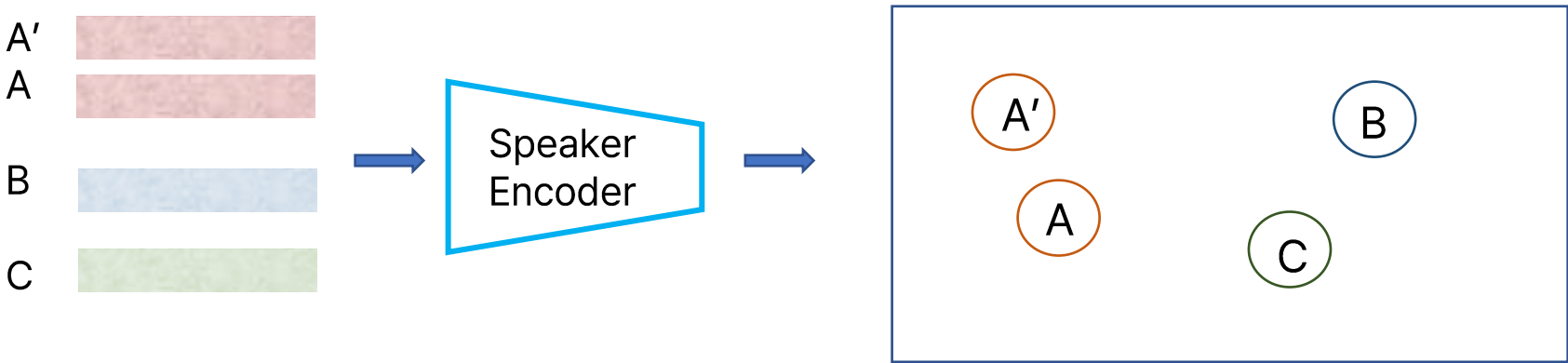


# Base Model

## Auto-VC

### Speaker Encoder 활용 결과

- ① 동일한 화자의 음성인 경우 비슷한 Speaker Feature를 도출할 수 있음
- ② 서로 다른 화자의 음성인 경우 Embedding 거리상 멀리 떨어져 있는 Speaker Feature를 도출.
- ③ 학습에 활용되지 않았던 음성도 Embedding 공간상에서 표현이 가능.

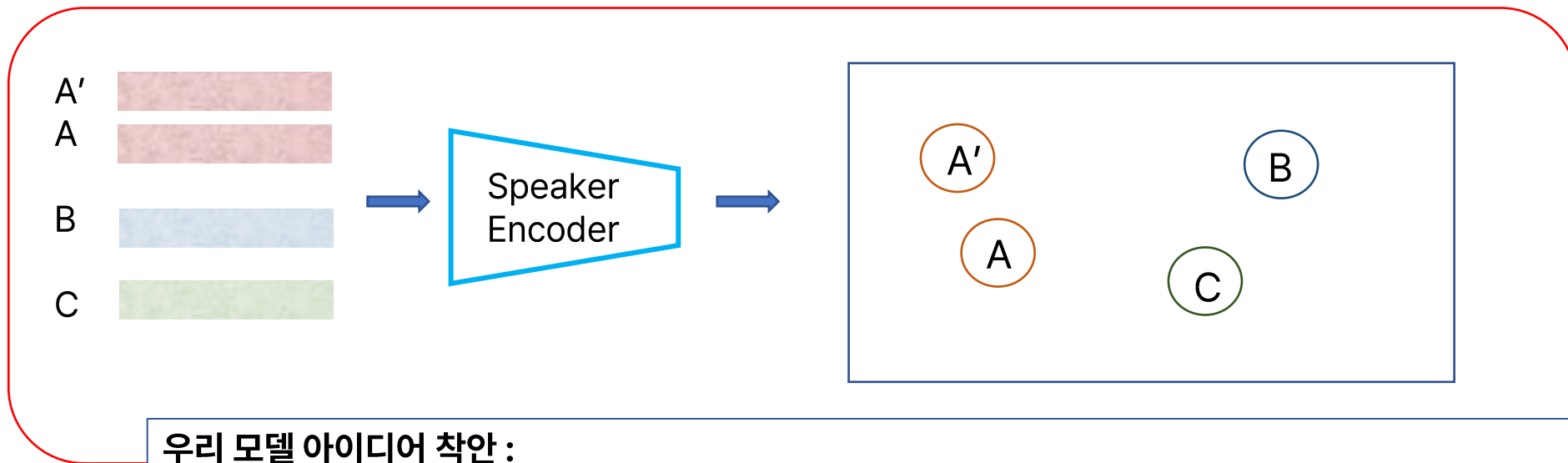


## Base Model

### Auto-VC

#### Speaker Encoder 활용 결과

- ① 동일한 화자의 음성인 경우 비슷한 Speaker Feature를 도출할 수 있음
- ② 서로 다른 화자의 음성인 경우 Embedding 거리상 멀리 떨어져 있는 Speaker Feature를 도출.
- ③ 학습에 활용되지 않았던 음성도 Embedding 공간상에서 표현이 가능.



**우리 모델 아이디어 착안 :**

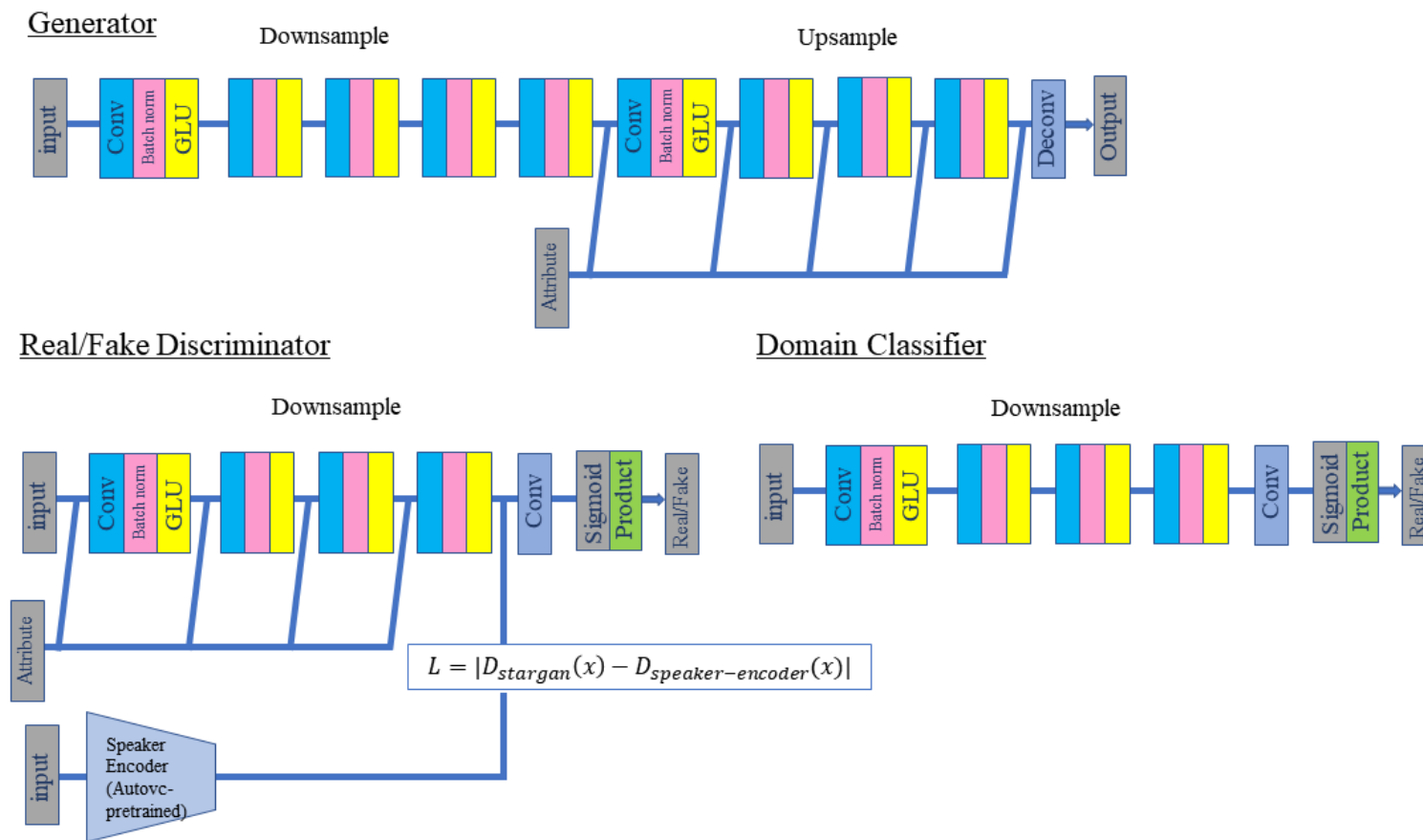
기존 StarGAN-VC 모델에 Speaker Encoder 모듈을 추가해 각 Speaker Feature 간 분류 정확도 향상을 기대



## Our Model

StarGAN-VC with speaker encoder

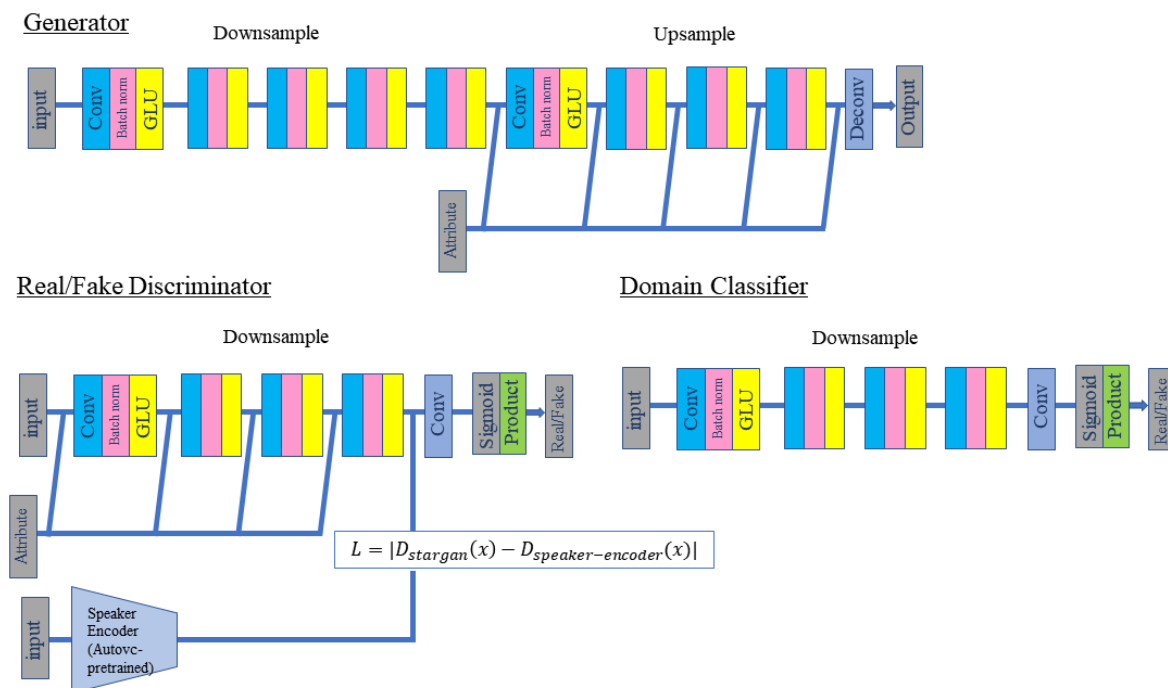
## Architecture



## Our Model

### StarGAN-VC with speaker encoder

## Architecture



- 기존의 StarGAN-VC 모델에 Auto-VC의 (pretrained 된) Speaker Encoder 모듈만을 가지고 옴
- StarGAN-VC 모델의 Discriminator 와 Generator 내 부에서 마지막 Convnet 들어가기 전에 Speaker Encoder에서 추출된 D-vector 과 Discriminator에서 추출된 D-vector 의 차이 값을 loss 값으로 정의해 학습을 진행

## Our Model

### StarGAN-VC with speaker encoder

#### Method

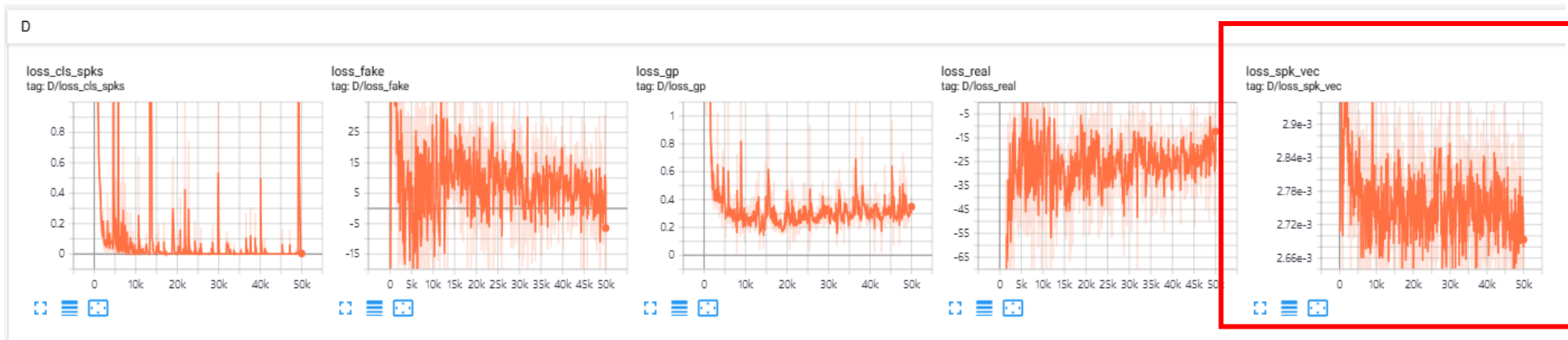
1. **Dimension 확인:** 기존 StarGAN-VC 모델에 loss function을 추가하기 위하여 Speaker Encoder 에서 추출되는 feature의 차원과 기존 VCTK의 데이터 차원을 맞춤 : VCTK 차원  $(x, 36) \rightarrow (x, 80)$
2. **Pre-training:** StarGAN-VC에서 진행되는 VTCK dataset으로 speaker embedding vector를 추출하는 pre-training 진행
3. **Discriminator 내부 loss 추가:** Discriminator class의 forward 함수에서, speaker embedding vector도 추가로 추출하도록 변환시키고, train 함수에서 변조된 음성 데이터의 embedding vector ( $x'$ )와 pre-trained ( $x$ )된 음성 데이터의 embedding vector를 Mean Square Error (MSE) loss로 비교하여 차이를 최소화하는 function 계산
4. **Generator 내부 loss 추가:** Discriminator과 동일하게 변조된 음성 데이터에서 추출한 embedding vector 와 pre-trained 된 음성 데이터의 MSE loss를 계산하여 차이를 최소화 시킴

## Results

### StarGAN-VC with speaker encoder

#### ❖ 모델 Loss 결과

##### • Discriminator



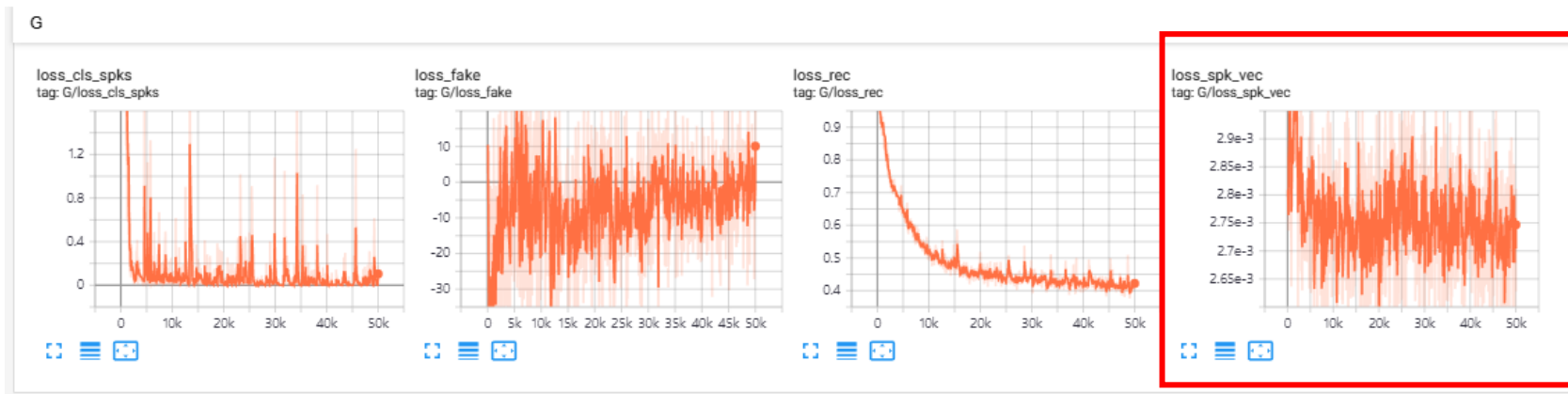
✓ 추가한 loss function에서 iteration횟수가 많아질수록 loss가 줄어드는 경향을 확인할 수 있음

## Results

### StarGAN-VC with speaker encoder

#### ❖ 모델 Loss 결과

##### • Generator



✓ 추가한 loss function에서 iteration횟수가 많아질수록 loss가 줄어드는 경향을 확인할 수 있음

Results

StarGAN-VC with speaker encoder

❖ 모델 평가방법: 정량적 지표

1. F0 Root Mean Square Error (RMSE) – p262 → p272 : iteration 50000

Content	AS-IS (StarGAN)	TO-BE (Our Model)
008	310.28	343.59
012	274.83	241.17
014	396.88	337.17
023	320.57	310.06
041	285.35	337.33
044	293.31	268.54
045	236.83	234.32
058	473.67	496.78
평균	323.97	321.12

Results

StarGAN-VC with speaker encoder

❖ 모델 평가방법: 정량적 지표



2. Mel-cepral Distortion (MCD) – p262 → p272 : iteration 50000

Content	AS-IS (StarGAN)	TO-BE (Our Model)
008	9.03	8.61
012	9.51	9.82
014	8.88	9.14
023	8.21	8.33
041	12.57	13.16
044	9.53	9.55
045	11.80	11.48
058	13.10	12.96
평균	10.33	10.38

Results

StarGAN-VC with speaker encoder

❖ 모델 음성 데이터 (P262 → P272 - content 012)

Original Voice	AS-IS (StarGAN)	TO-BE (Our Model)
		



## Conclusion

### StarGAN-VC with speaker encoder

#### ❖ 연구의 한계

- ✓ **[훈련 시 iteration 수의 부족]** 기존 연구들이 훈련 시의 iteration 을 200000 번 수행한 것에 비해, 우리 연구는 시간 관계 상 50000 번의 훈련 iteration을 시행했다. 기존 연구에 비해 적은 iteration 을 수행한 결과 값이기 때문에 추후 기존 연구와 동일한 학습 iteration 수행을 통해 성능을 높일 수 있을 것이라 기대할 수 있다.
- ✓ **[손실함수의 다양성 부족]** 현재 모델의 loss 값은 Discriminator 에서 추출한 embedding vector 과 pre-trained 된 음성 데이터 간의 MSE loss 를 계산하여 차이를 최소화 시키는 방향으로 구현되어 있다. 선행 연구에 따르면<sup>1)</sup>, MSE loss 는 모델 성능을 높이는 것에 한계가 있어 SI-SNR 등 다양한 손실 함수를 통해 성능을 높이는 연구가 진행되고 있다. 추후 보다 다양한 손실 함수를 통해 성능 개선이 가능할 것이라 기대할 수 있다
- ✓ **[source – target 조합의 부족]** 기존 starGAN-VC 에서는 다양한 source speaker-target speaker 조합을 통해 결과 샘플을 도출하였으나 시간 관계상 p262, p272 두 명의 발화자만 추출해 결과 샘플을 추출하였다. 추후 다양한 조합을 통한 결과 도출을 실험해 볼 수 있을 것이라 기대할 수 있다.

1) 황서림, 변준, 박영철 (2021), 다양한 손실함수를 이용한 음성 향상 성능 비교 평가, 한국음향학회지

## Conclusion

### StarGAN-VC with speaker encoder

#### ❖ 결론

- 우리가 연구를 진행한 StarGAN-VC with speaker encoder 모델은 기존의 StarGAN-VC 모델에 Auto-VC 모델의 speaker encoder 모듈을 추가하여 보다 정확한 voice conversion 이 가능하도록 구현한 모델입니다.
- RMSE 와 MCD, 두 가지 정량적 지표를 통해 기존 StarGAN-VC 모델과 우리 모델을 비교했을 때, 평균적으로 RMSE 의 경우 우리 모델이 더 낮았고, MCD 의 경우 소폭 StarGAN-VC 모델이 높게 도출되었습니다. 명확한 성능 개선을 보였다고 말하기는 어려우나, 평균적으로 RMSE 값이 우리 모델이 더 낮았다는 점, MCD 값 또한 기존 StarGAN-VC 모델의 MCD 값에서 크게 벗어나지 않았다는 점에서 추후 연구에서 개선을 통해 명확한 성능 개선을 보일 것이라 기대할 수 있습니다.
- 우리 모델은 크게 훈련 iteration 의 부족, 손실함수의 다양성 부족, source speaker 과 target speaker 간의 조합 부족이라는 세 가지 한계점을 가지고 있습니다. 시간 제약으로 인한 한계점이었기 때문에 추후 연구에서 해당 부분에 대한 개선을 통해 보다 정확한 voice conversion을 수행할 수 있을 것이라 기대할 수 있습니다.