야구선수의 응원가 쟁탈기 

Audio

용서규합니다 김규리 김동규 서은서 이재용

1

Motivation

야구선수가 응원가를 “직접” 부른다면?

야구, 좋아하시나요?

'최강야구'·'찐팬구역'·'야구대표자'...야구 예능 전성시대

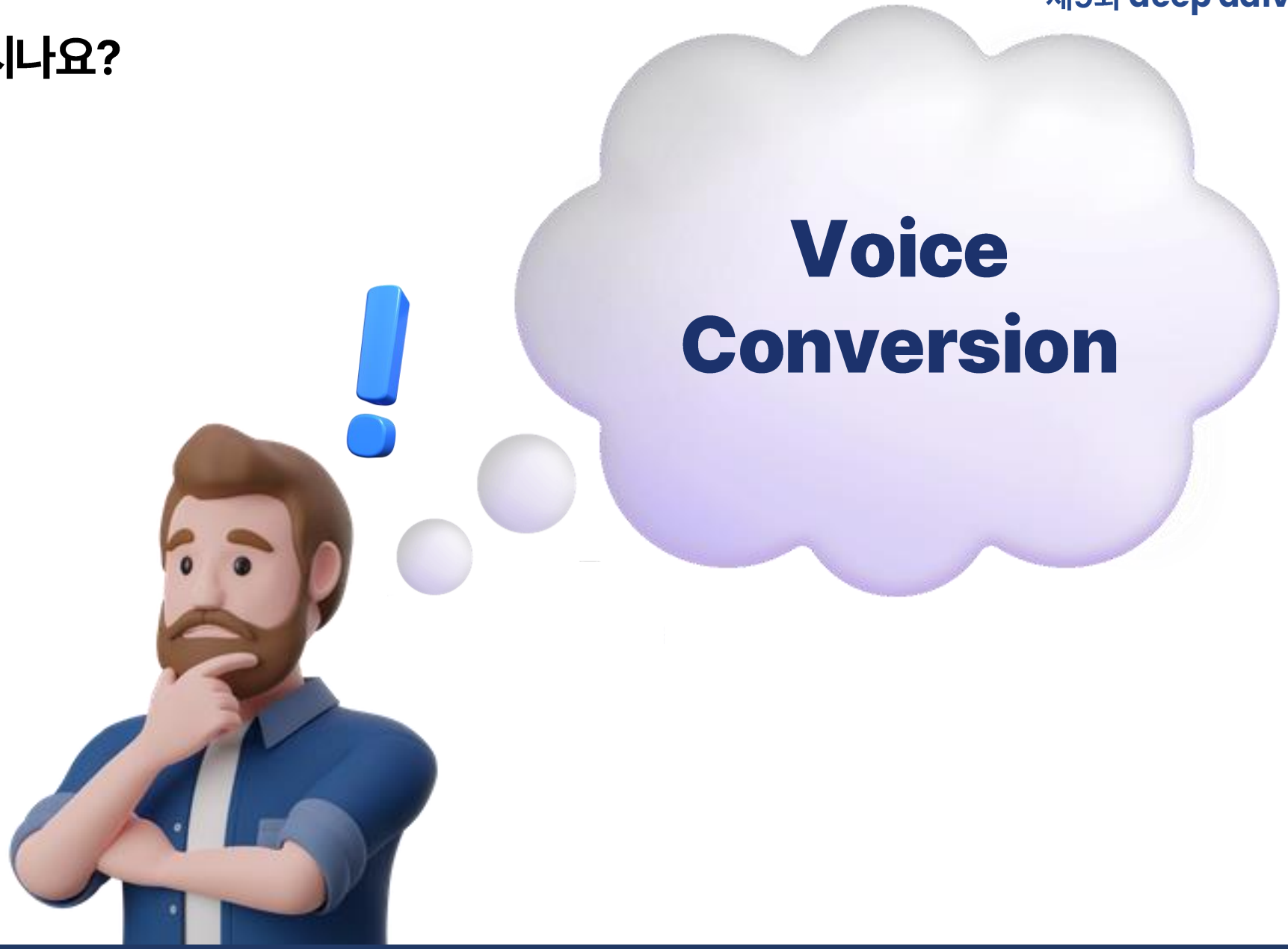
-한국경제-

2024 KBO 응원가 신곡들 차트 진입?! 당신의 원픽은?

-엠빅뉴스-

야구, 좋아하시나요?

야구, 좋아하시나요?



Voice Conversion

진실은 언제나 하나!

Voice Conversion

진실은 언제나 하나!

Methodology

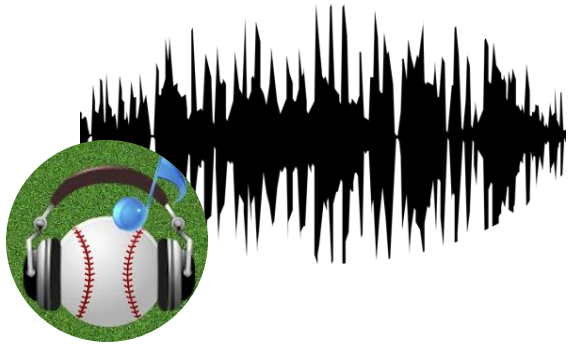
AutoVC

2

Methodology

Pipeline

선수 목소리 wav
(speaker data)

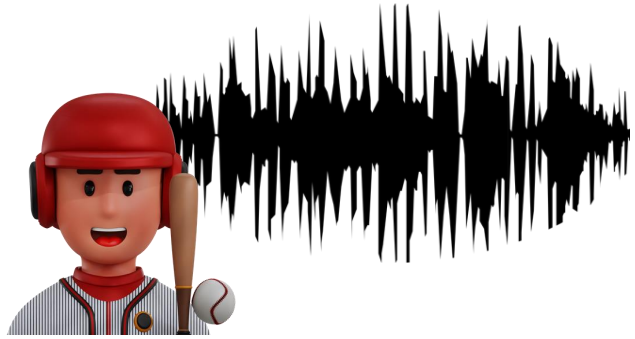


야구 응원가 wav
(content data)

Input

AutoVC

Output



선수가 부르는 응원가 wav

Methodology

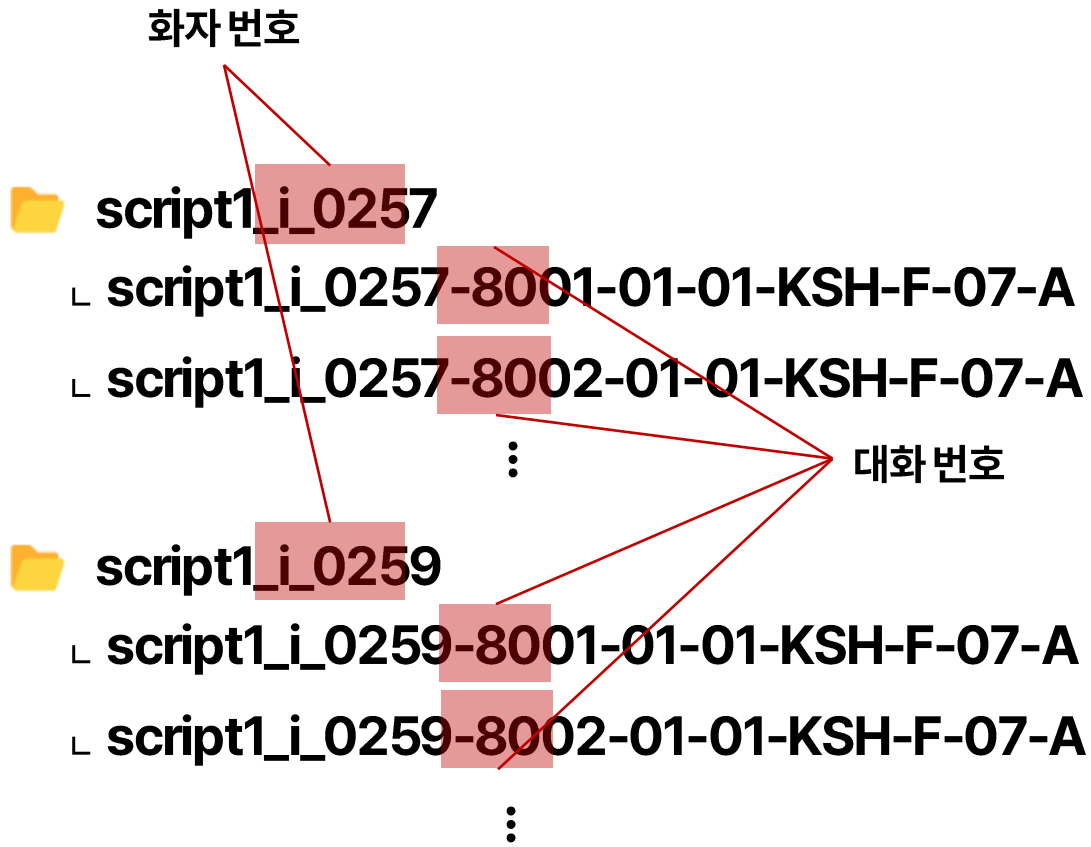
Train Data

AI Hub

명령어 음성(일반 남여)

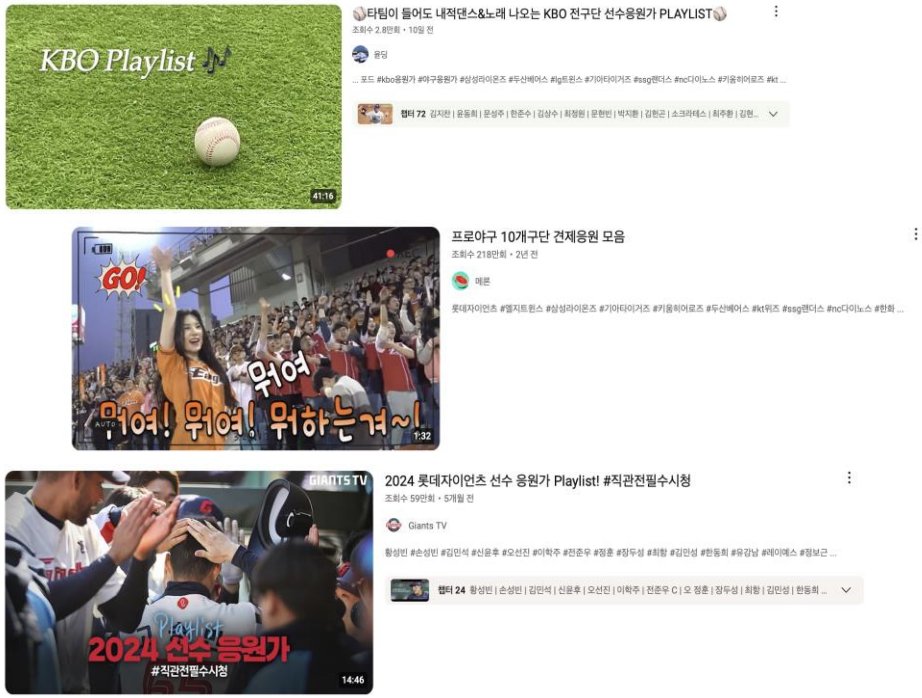
분야 한국어

유형 오디오, 텍스트



Methodology

Conversion Data - Source



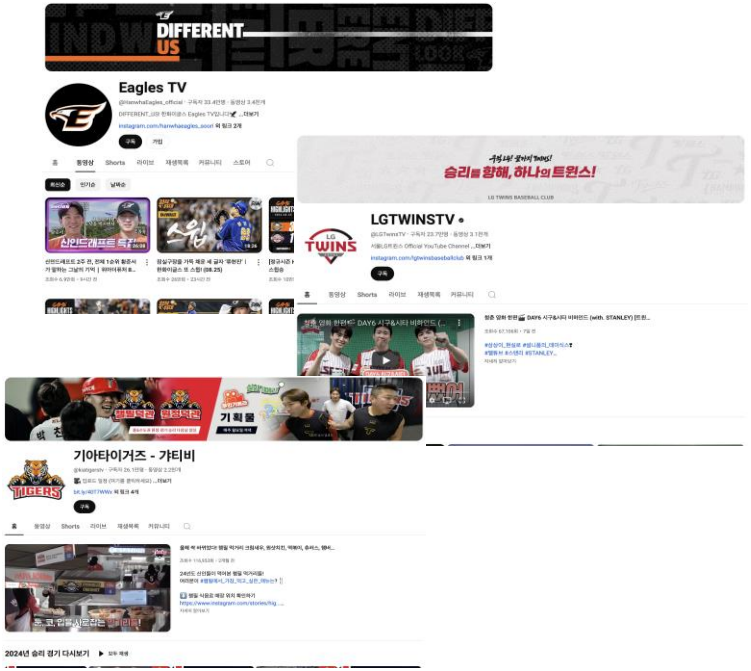
응원가 wav

source separation

보컬 추출 wav

Methodology

Conversion Data - Target



source separation

선수 인터뷰 wav

보컬 추출 wav

Methodology

Motivation

기존 VC Systems

Parallel Speaker Data

Many-to-Many VC 연구의 미흡

Zero-shot 불가

+) GAN과 CVAE 간의 trade-off
(Conditional Variational AutoEncoder)

AutoVC

Non-Parallel Data로 Many-to-Many VC 가능

Zero-shot 가능

+) Style Transfer의 도입

+) 간단한 Auto-Encoder 프레임워크 기반

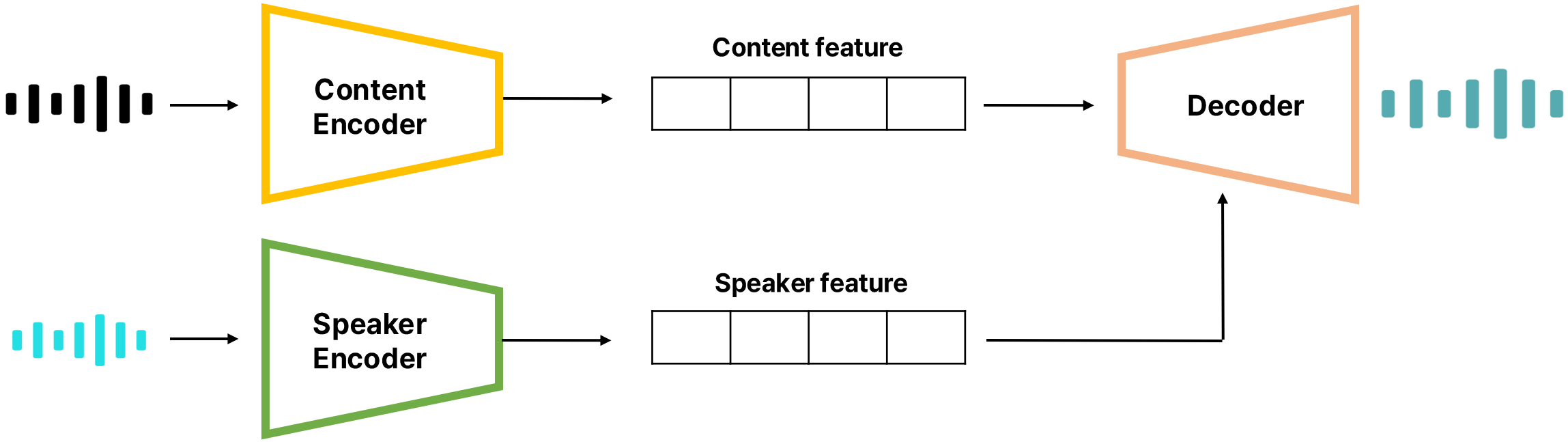
Methodology

Architecture Overview

① **Content Encoder** 음성 Feature로부터 Content Embedding 생성 모듈

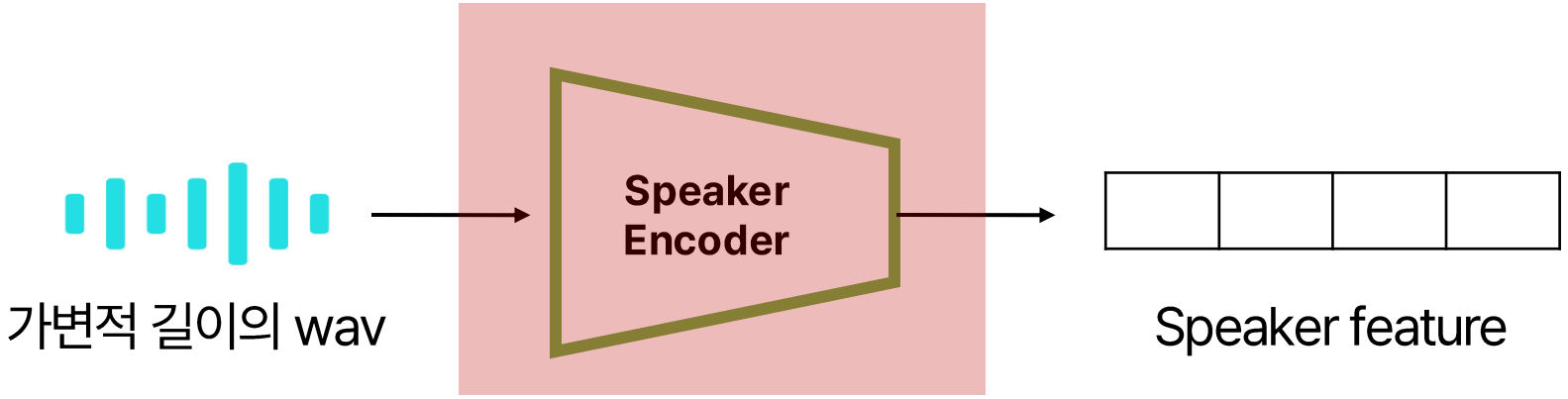
② **Speaker Encoder** 음성 Feature로부터 Speaker Embedding 생성 모듈

③ **Decoder** Content Embedding과 Speaker Embedding을 입력받아 Speaker 특징이 포함된 Speech 생성하는 모듈



Methodology

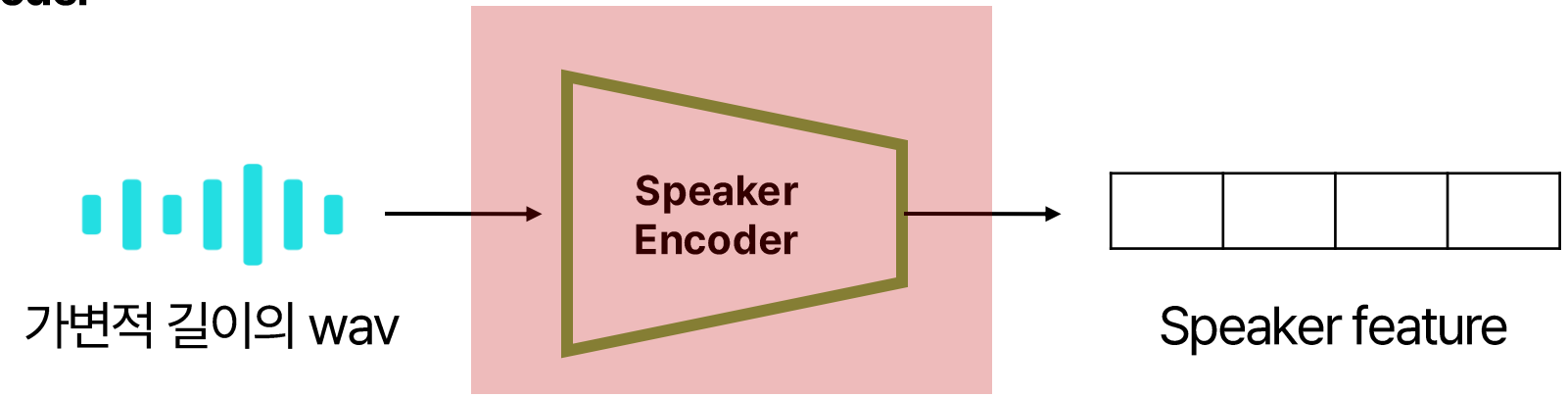
1. Speaker Encoder



① 일관성 ② 유사성

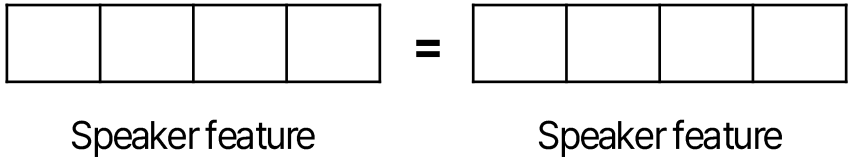
Methodology

1. Speaker Encoder



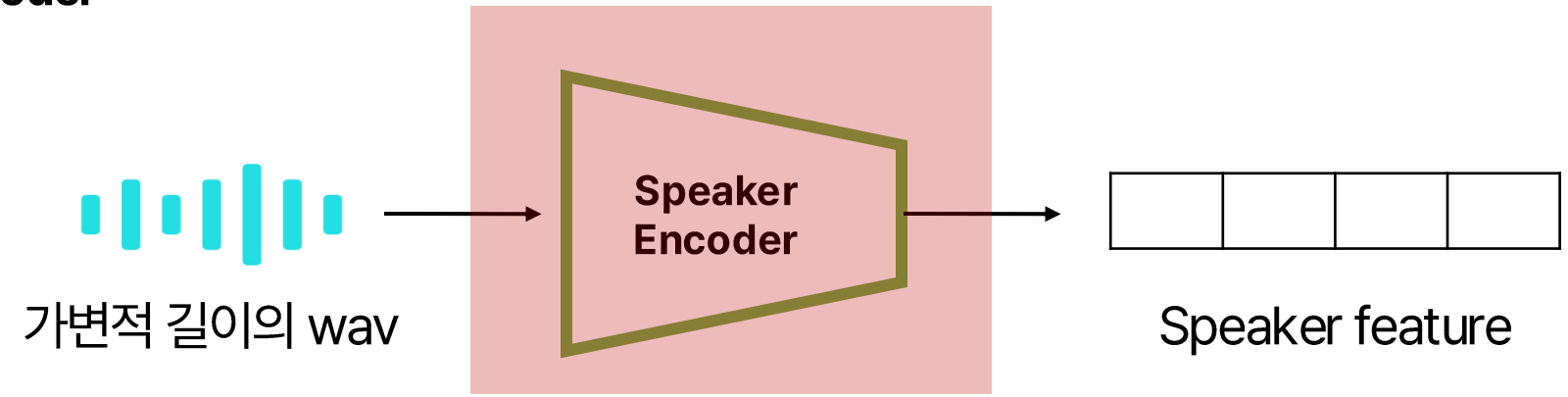
① **일관성** ② **유사성**

동일한 화자 음성 wav



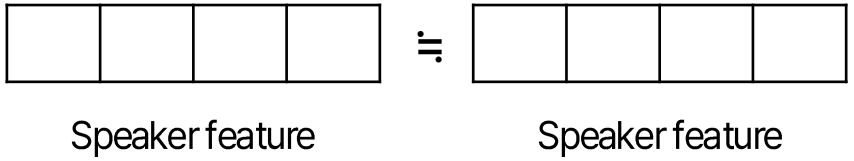
Methodology

1. Speaker Encoder



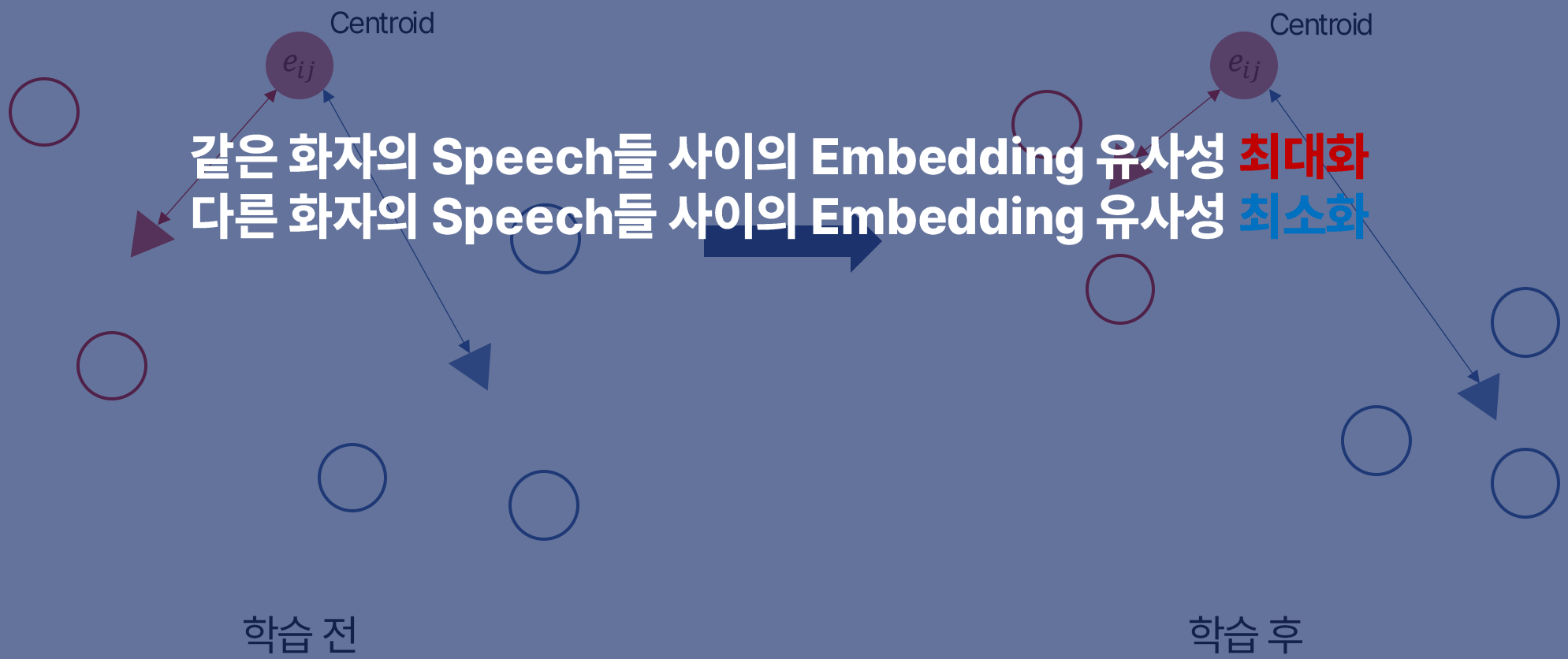
① 일관성 ② 유사성

비슷한 음성의 화자 wav



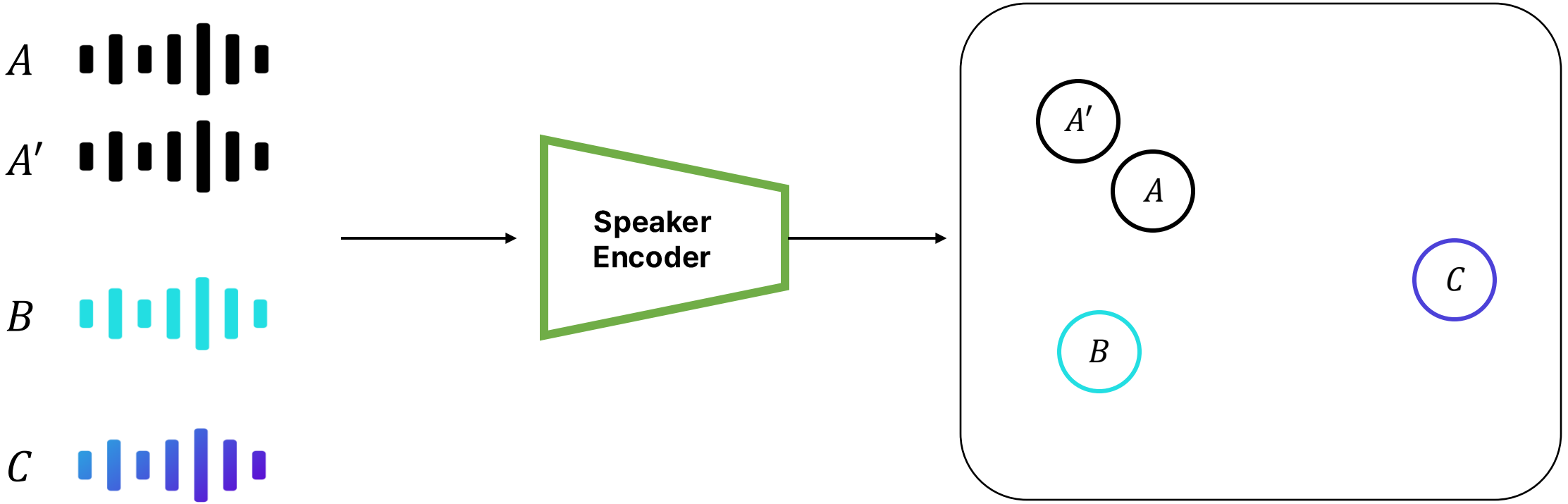
Methodology

1. Speaker Encoder – 학습법



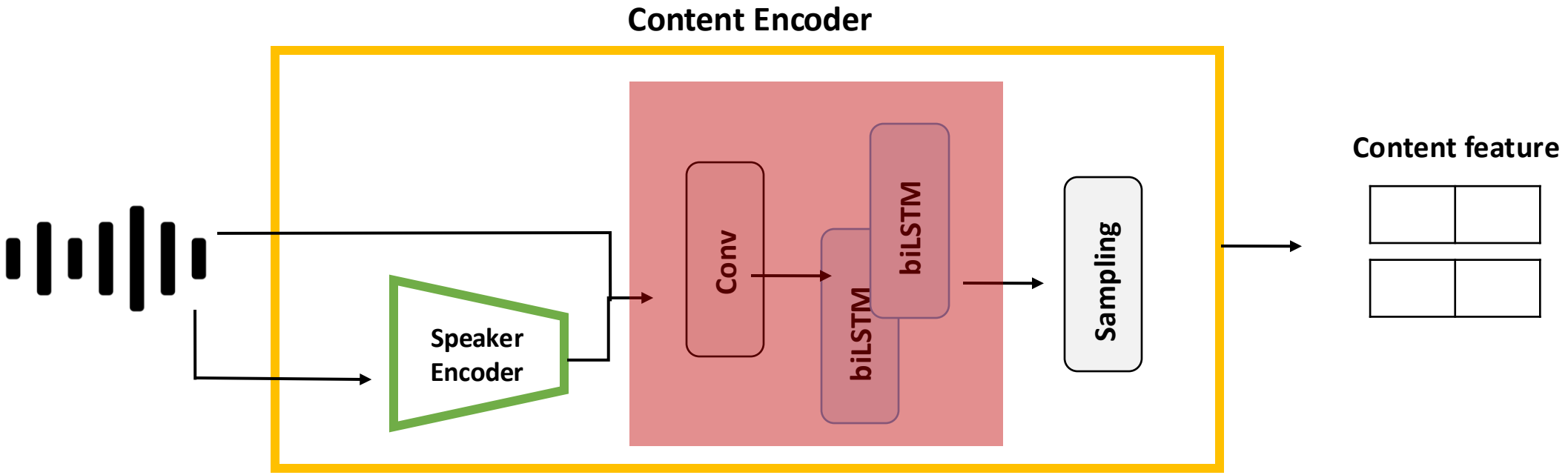
Methodology

1. Speaker Encoder



Methodology

2. Content Encoder



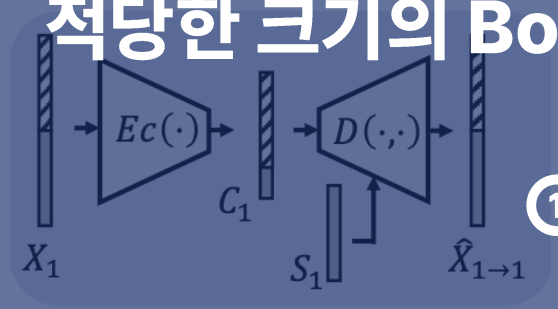
BottleNeck 구조!

Methodology

2. Content Encoder - BottleNeck

적당한 크기의 BottleNeck으로 학습됐을 때 Content Encoder는

① 음성에서 Speaker 정보를 배재,
② Content 정보만 선택,
③ Decoder를 통한 좋은 품질의 음성을 생성

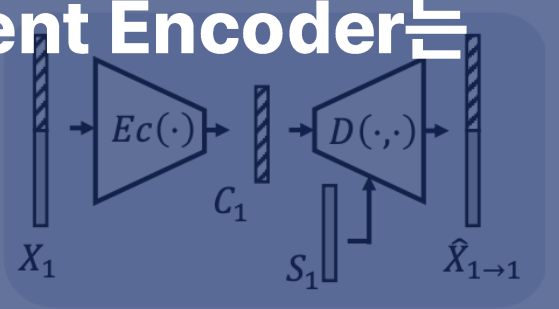


(a) Bottleneck too wide

Content Embedding에
Speaker 정보 포함

(b) Bottleneck too narrow

Content Embedding에
Content 정보 누락

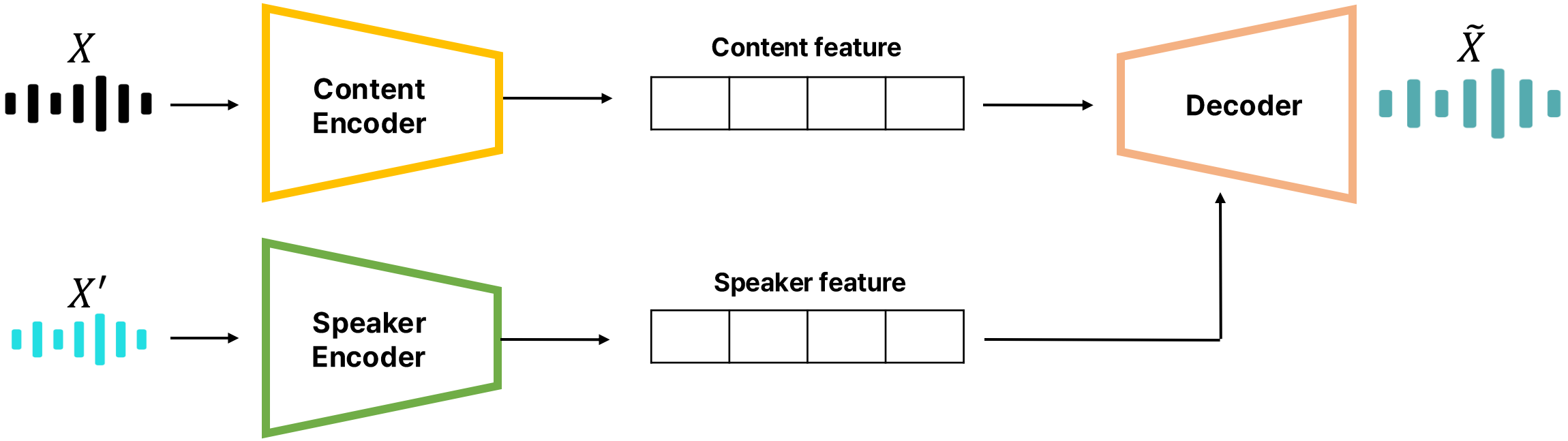


(c) Bottleneck just right

Content Embedding에
Content 정보만 포함

Methodology

3. Decoder



$$L_{recon0} = \mathbb{E}[\| \tilde{X}_{1\to1} - X_1 \|_2^2]$$

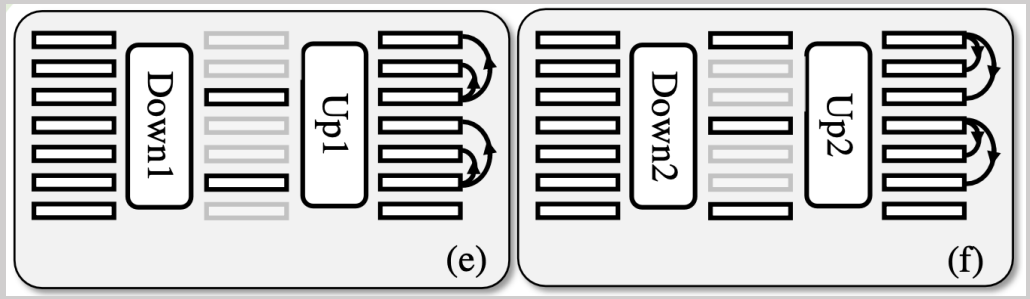
Experiments

3

Experiments

개선점 1 - Upsampling & Downsampling

기존 AutoVC의 Sampling method



선택된 일정 간격의 샘플들에 의존적

▼

나머지 샘플들의 특징 반영 불가

개선된 Sampling method

Downsample Operator

Average Pooling & 1D-Convolution (big strides)

+

Upsample Operator

Transposed Convolution, Interpolation

▼

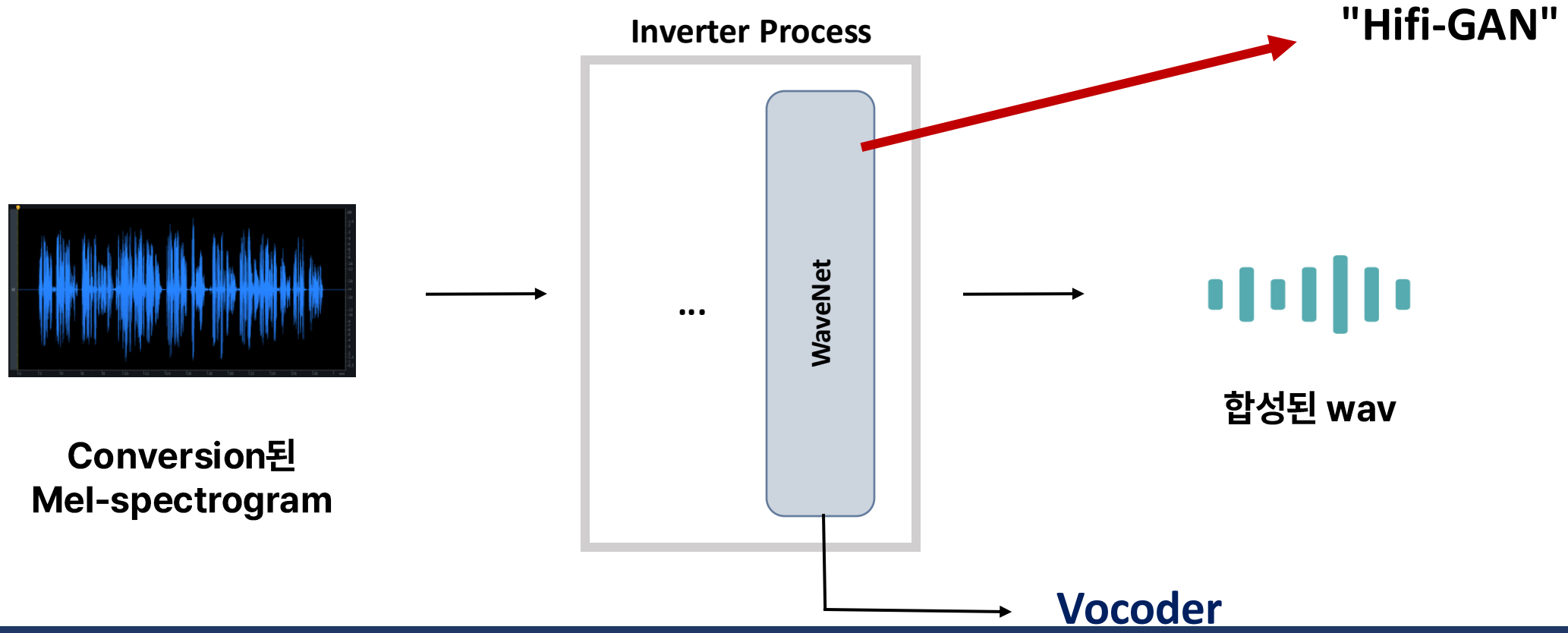
최적의 조합 찾기!

Experiments

개선점 2 - Hifi-GAN으로 Vocoder 대체

- 기존 Vocoder로 사용되었던 WaveNet (2016)을 Hifi-GAN (2021)으로 변경

▶ 고품질의 audio와 빠른 합성 속도를 기대



Results

4

Result

- VC의 결과가 잘 되지 않은 것을 토대로 다양한 관점에서 실패 요인 탐색 진행

1. AutoVC에서 한국어 학습의 어려움
    - Vocoder와 AutoVC에서의 분포를 일치시키는 것에 어려움이 존재

2. 오디오 도메인에서의 학습의 어려움
    - 화자 1명 당 최소 1시간의 데이터가 필요
    - 모델의 구조가 달라지는 경우, 사전 학습된 데이터를 사용하지 못하는 경우가 존재
    - 즉각적인 피드백, 확인이 힘듦

감사합니다

Q&A