박송은 백광제 신현경 임진혁 전민규
정윤호

TOBIG`S Rhapsody

tacotron을 이용한 음성 합성기 제작

아이유

said

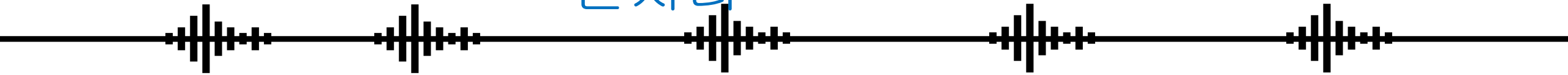
민규야! 안녕! 너와 친구가 되고 싶어!

INDEX

주제 선정

데이터 전처리

결론 및 제언



데이터 준비

모델 구조

1 주제 선정

음성 합성이란?

인위적으로 사람의 목소리를 합성하는 시스템이며

텍스트를 음성으로 변환하는 기술로 **Text-To Speech**

줄여서 **‘TTS’**라고도 한다.

1 주제 선정

음성합성 도전!

▼

좋아하는 연예인의 목소리를 만들어서 활용해보고

싶다!

▼

질 높은 데이터가 아닌 적은 양의 비정형 데이터로도

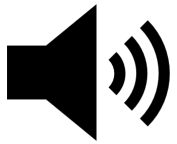
음성 합성이 어느정도 까지 가능한지 알고 싶다!

1 주제
선정

질 높은 [redacted] 는 적은 데이터로도

음성 [redacted] 정도 까지 알고 싶다!

2 데이터



모델 학습을 위해 **음성 데이터와 스크립트 데이터 쌍**이 필요

-> **KSS** 데이터 셋과 유튜브(**STT**스크립트) 사용

2 데이터

KSS DATASET

a Korean single speaker speech dataset

IU DATASET

Youtube링크와 타사이트에서 가져온 아이유의 speech dataset

TAEYON DATASET

Youtube링크에서 가져온 태연의 speech dataset

2 데이터 종류

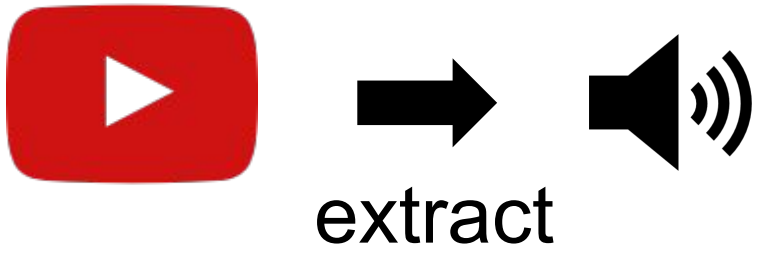
IU DATASET

Youtube링크와 타사이트에서 가져온 아이유의 speech dataset

TAEYON DATASET

Youtube링크에서 가져온 태연의 speech dataset

2 데이터 준비



pytube : 유튜브에서 동영상 추출

moviepy : 동영상을 음성 파일로 변환

2 데이터 준비

PROBLE

M

일단 유튜브 영상에서 음성 파일을

가져왔는데

어떻게 전처리를 할까?

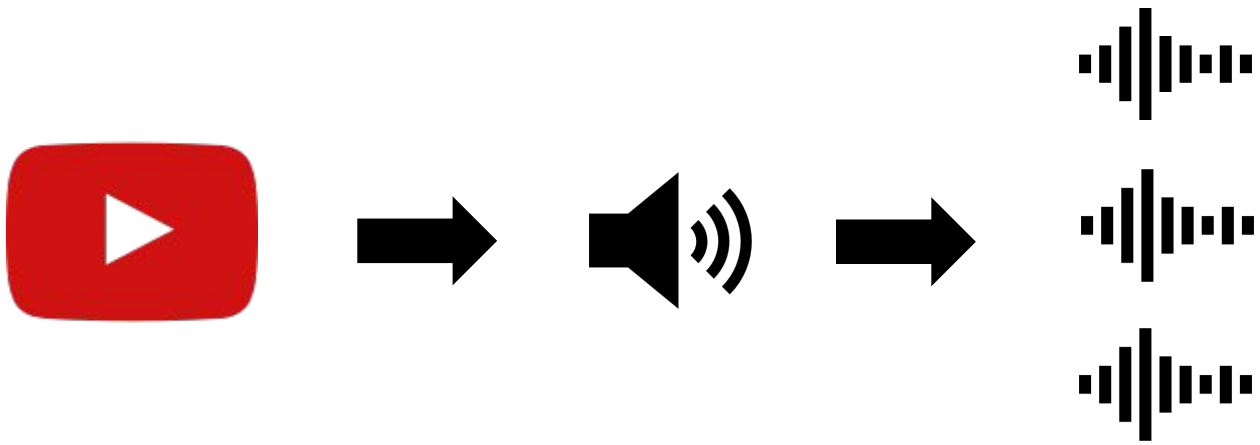
2 데이터 준비

PROBLEM

1. 음성 데이터의 분할(Segmentation)?

2. 음성 데이터와 쌍을 이룰 스크립트 데이터 생성?

3. 최소 학습량을 확보?

2 데이터

segmentation

2 데이터 준비

기준

- 10초 단위로 분할 -> 음성파일 1900개 / 5시간 확보

segmentation

2 데이터 중심

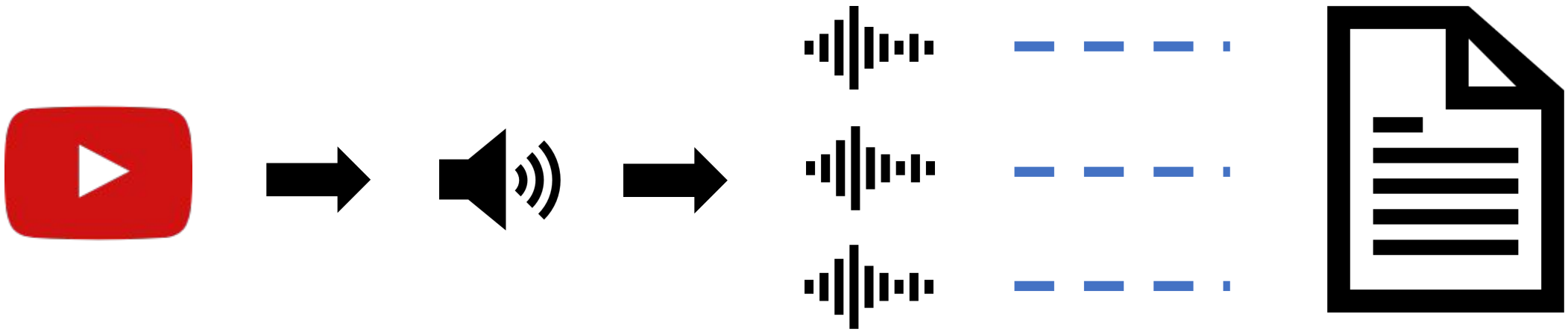
segmentation 기준

- **10초 단위로 분할**
- > 분할 포인트가 중구난방
- > 다른 사람 목소리, 배경소음, 노래, 긴 침묵 등 다양한 변수
- > 처음과 시작이 불분명

2 데이터 준비

segmentation 새로운 기준

- **무음 단위로 분할**
- 데시벨을 조절해 낮은 소리는 제거
- 2초 이상 12초 이하의 파일만 사용

2 데이터


STT(Speech To Text)

2 데이터 종류

STT

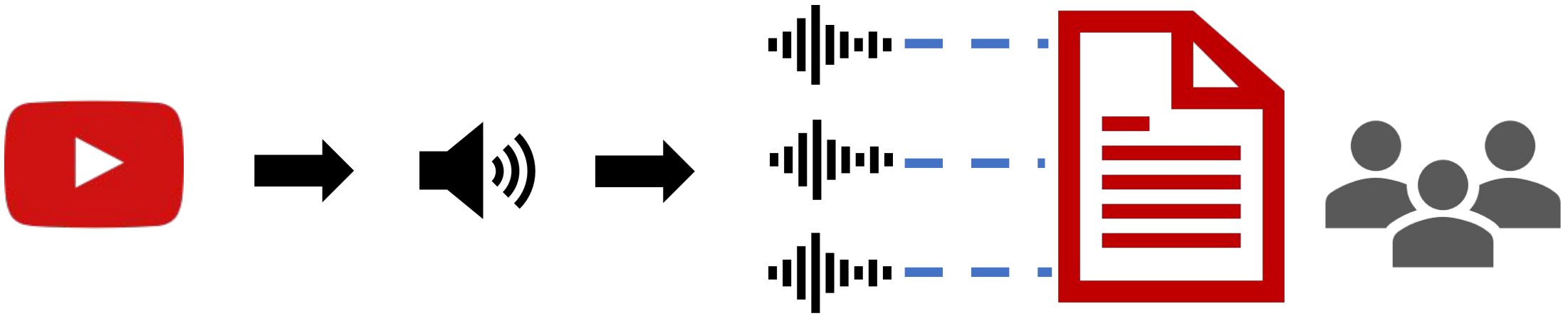
?

- **S**peech **T**o **T**ext (음성 -> 텍스트)
- Google Cloud Speech API : 음성 데이터에 사용할 수 있는 머신러닝 기반 **텍스트 변환 API**

유튜브 자막, 구글 어시스턴트에 적용되고 있는,

구글 클라우드 서비스가 제공하는 “받아쓰기 기능 "

2 데이터 중비



STT 결과 정제(수작업)

2 데이터 정제

STT 결과

정제 규칙

1 표준어 문법을 원칙으로 하되, 최대한 발음을 살리는 방향으로 표기

2 영어 -> 한글

3 ,(쉼표) ?(물음표) .(마침표)

9_9_0000.wav	그런 거 아니거든요 자존심 때문에 전화 안 한다고 또
9_9_0001.wav	제일 다 해 가지고
9_9_0002.wav	현재 위치에 고객님 많이 벌었어요이 지금도 재희 고기고
9_9_0003.wav	잊어야 한다는 마음으로 편곡도 재밌지가 해 줬어 밤편지도 1시간
9_9_0004.wav	동생들이
9_9_0005.wav	응원합니다 선배님과 재희 분한테 고마운 마음을 담아서
9_9_0006.wav	또 다른 우리 팀들 저기 친구들
9_9_0007.wav	머리 긴 여자 분 계세요
9_9_0008.wav	error
9_9_0009.wav	이거 눌러 버렸네 우리가
9_9_0010.wav	그러면 마지막 마음먹은 될까요
9_9_0011.wav	몇 시에요 진짜로 8시 50분입니다 여러분
9_9_0012.wav	거야 진짜로 가지고
9_9_0013.wav	가세요
9_9_0014.wav	예
9_9_0015.wav	error
9_9_0016.wav	error
9_9_0017.wav	그래서
9_9_0018.wav	행운이 보고 참 좋은 일이라고 생각합니다
9_9_0019.wav	error
9_9_0020.wav	error

2 데이터 준비

STT 결과

정제

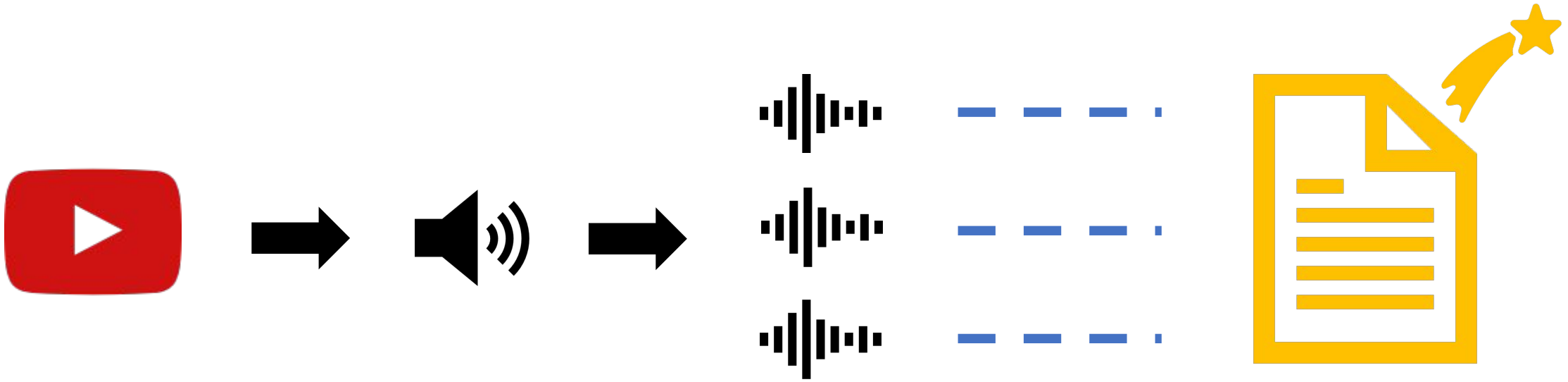
-> 정제 작업 후 삭제한 데이터가 증가해 학습 데이터가 부족해짐

-> 10초 분할 데이터 중 양호한 데이터 중복 사용

총 1229개 / 156분 음성 데이터 확보!

9_9_0000.wav	그런 거 아니거든요 자존심 때문에 전화 안 한다고 또
9_9_0001.wav	제일 다 해 가지고
9_9_0002.wav	현재 위치에 고객님 많이 벌었어요이 지금도 재희 고기고
9_9_0003.wav	잊어야 한다는 마음으로 편곡도 재밌지가 해 줬어 밤편지도 1시간
9_9_0004.wav	동생들이
9_9_0005.wav	응원합니다 선배님과 재희 분한테 고마운 마음을 담아서
9_9_0006.wav	또 다른 우리 팀들 저기 친구들
9_9_0007.wav	머리 긴 여자 분 계세요
9_9_0008.wav	error
9_9_0009.wav	이거 눌러 버렸네 우리가
9_9_0010.wav	그러면 마지막 마음먹은 될까요
9_9_0011.wav	몇 시에요 진짜로 8시 50분입니다 여러분
9_9_0012.wav	거야 진짜로 가지고
9_9_0013.wav	가세요
9_9_0014.wav	예
9_9_0015.wav	error
9_9_0016.wav	error
9_9_0017.wav	그래서
9_9_0018.wav	행운이 보고 참 좋은 일이라고 생각합니다
9_9_0019.wav	error
9_9_0020.wav	error

2 데이터 준비



데이터셋 완성!

2 데이터 준비

KSS DATASET

IU DATASET

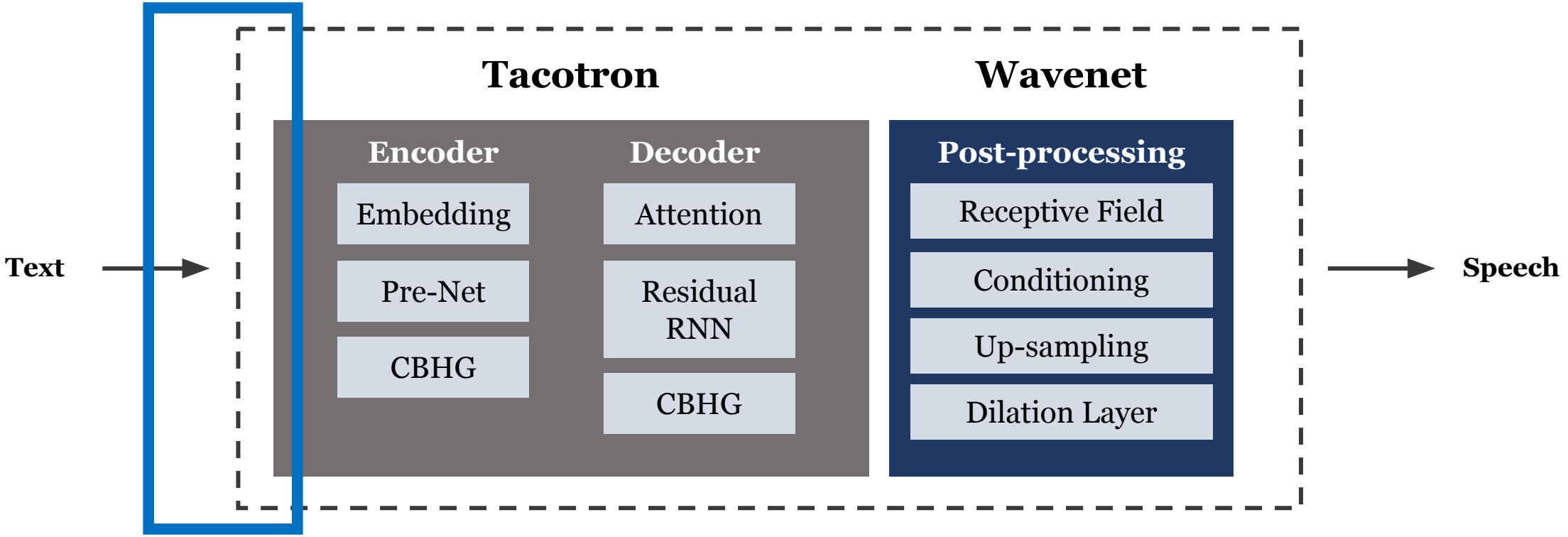
TAEYON DATASET

KSS 6시간 30분 + 아이유 2시간 36분 + 태연 2시간 2분

확보!!

3 데이터 전처리

음성 합성 프로세스 모델링



모델 입력단에 들어가기 위한 전처리 필요!

3 데이터 전처리

- 한글 string을 초,중,종성으로 나누고 마지막에 end token('~') 붙임

  [안녕] => [ㅇㅏㄴㄴㅕㅇ] 으로 쪼개고 숫자로 매칭

- 영어 / 발성법이 여러가지인건 미리 사전에 정의

  dict)

  DJ : 디제이

  2018 : 이천십구

- 전처리 결과 : audio, mel, linear npy 파일 + 텍스트

3 데이터 전처리

1. audio.npy

- wav파일

- librosa.core.load(sr=24000)[0]으로 1차원 wav파일 추출

- [-1,1]로 resacling

3 데이터 전처리

2. mel spectrogram

1) wav 파일을 preemphasis로 잡음제거 후 stft

2) librosa.filters.mel과 1)의 절댓값을 dot product

3) 2)의 amplitude spectrogram에 대해 소리 크기도 비선형성을 나타내기 위해 **log**를 취하여 **dB-scaled spectrogram**으로 바꿈

4) 3)에서 ref_level_db를 빼고 normalize

3 데이터 전처리

3. Linear spectrogram

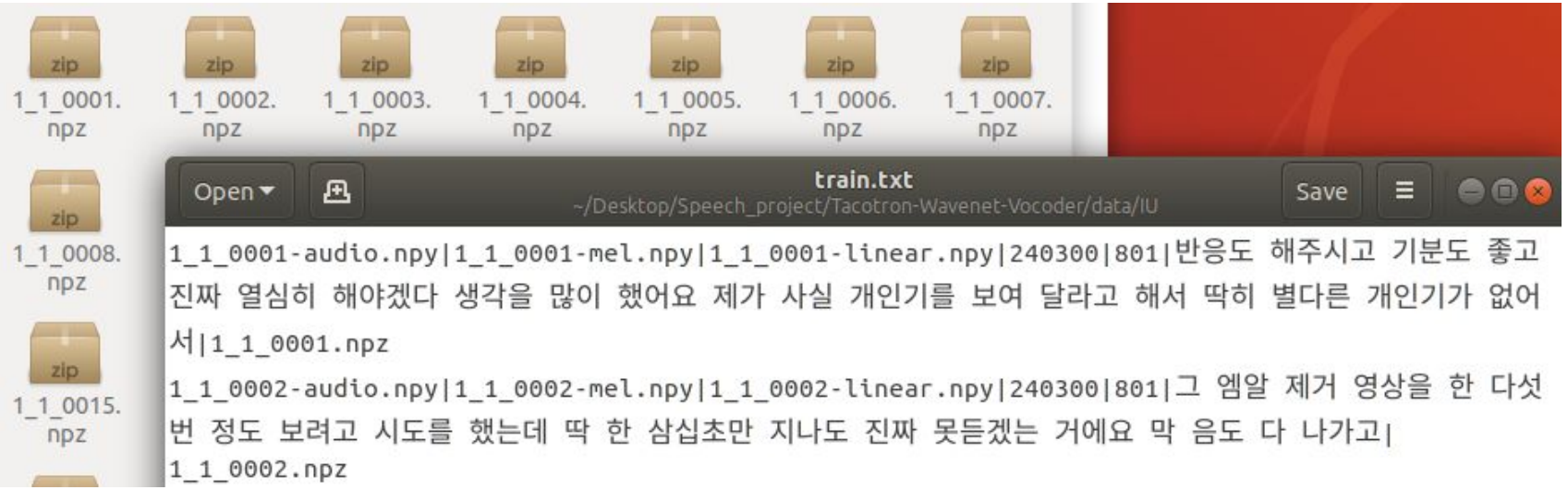
1) **mel과 동일**

2) 1)결과에 절댓값을 씌운 후 dB scaled Spectrogram으로 변환

3) 2) 결과에 ref_level_db를 빼고 normalize

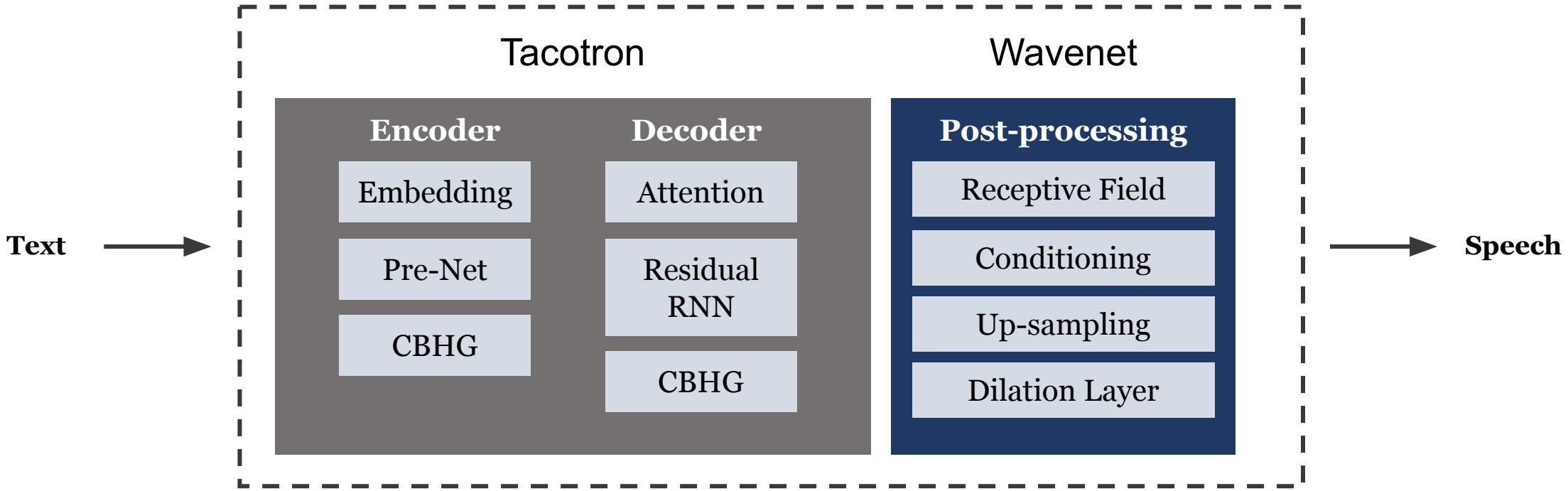
3 데이터 전처리

전처리 결과물 : audio, mel, linear npy 파일 + 텍스트

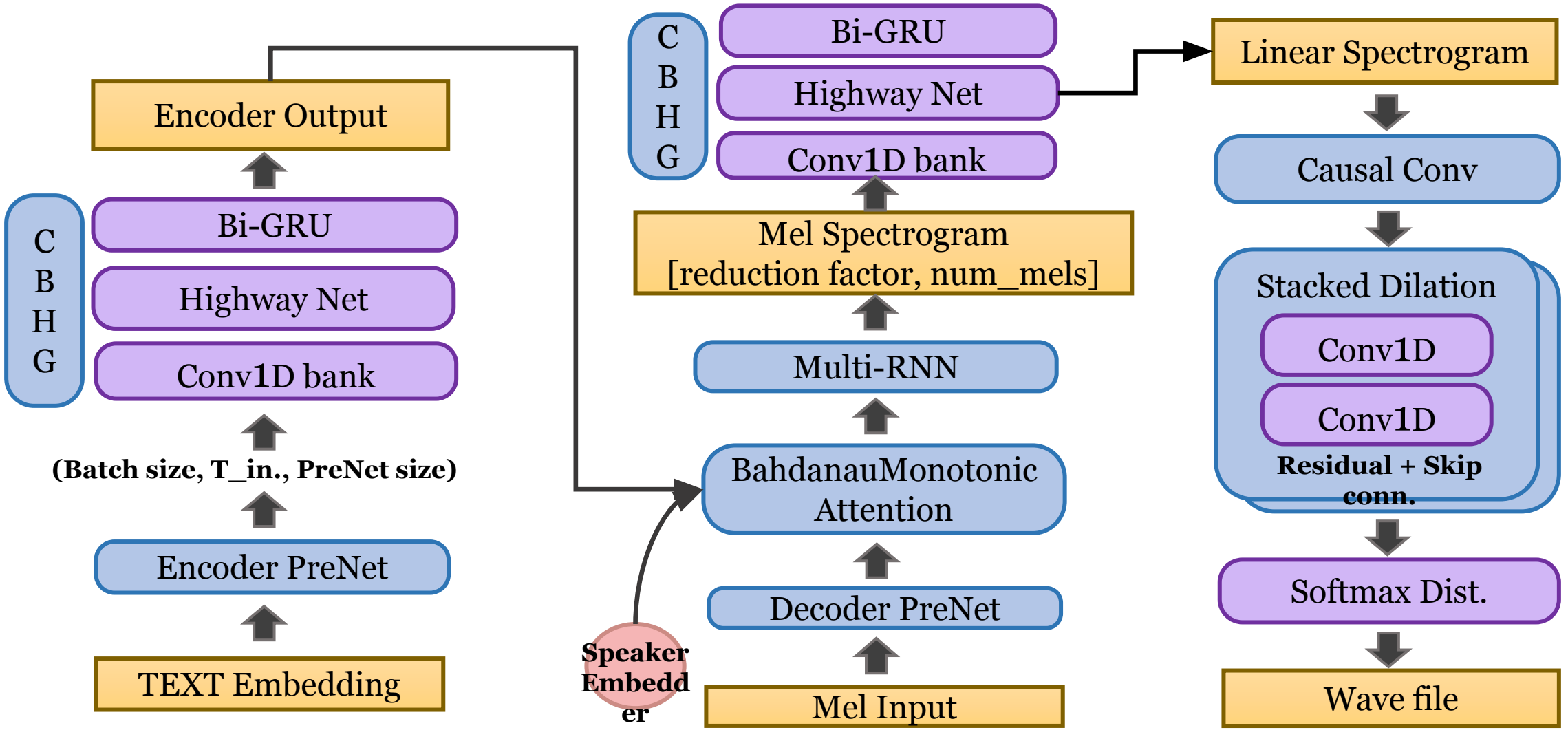


4 모델 구조

음성 합성 프로세스 모델링



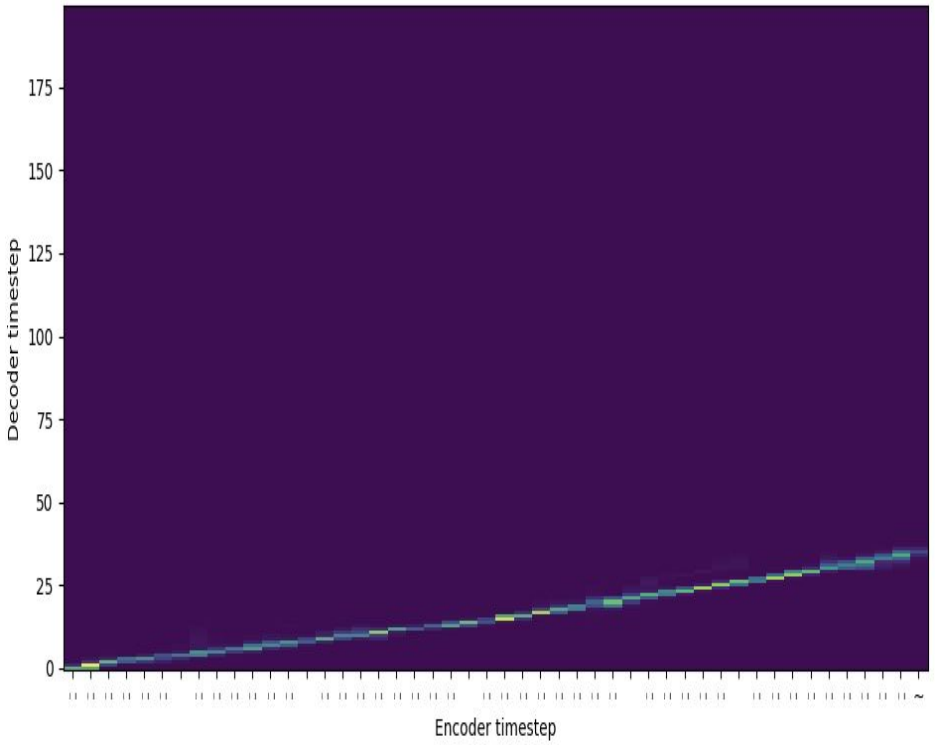
4 모델 구조


Encoder Output
C
B
H
G
Bi-GRU
Highway Net
Conv1D bank
(Batch size, T_in., PreNet size)
Encoder PreNet
TEXT Embedding
C
B
H
G
Bi-GRU
Highway Net
Conv1D bank
Mel Spectrogram
[reduction factor, num_mels]
Multi-RNN
BahdanauMonotonic Attention
Decoder PreNet
Mel Input
Speaker Embedder
Linear Spectrogram
Causal Conv
Stacked Dilation
Conv1D
Conv1D
Residual + Skip conn.
Softmax Dist.
Wave file

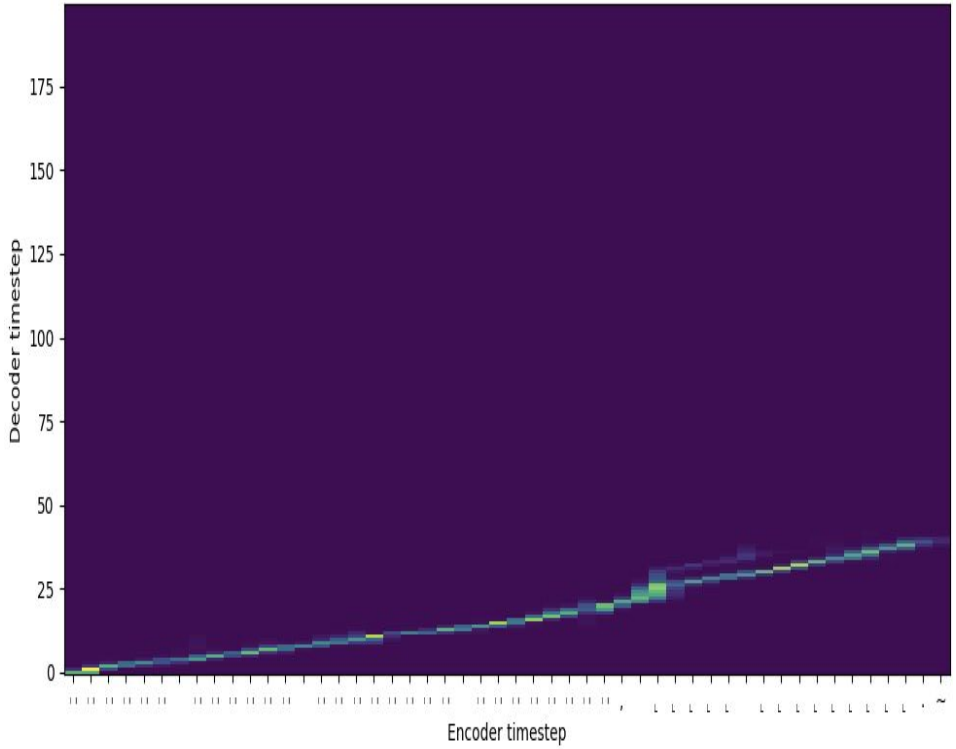
5 결론

그래프 분석

띄어쓰기 경우



‘,’ 경우



5
결론

합성

결과

5
결론

활용방

안

1. 문자를 읽기 어려운 장애인에게 기계로 하고 싶은 말을 쓰면 자신이 원하는 목소리를 선택해 목소리를 '대신' 내줄 수 있는 시스템 제작
2. 독거 노인들에게 가족 등 친숙한 목소리로 응답할 수 있는 음성 AI 스피커 제작
3. 말로 발표하는 것이 곤란한 사람이 스피치 대체 수단으로서 이용할 수 있음

5
제언

- **데이터 측면**

비정형 데이터 특성상 버려지는 데이터가 많다는 문제

배경음 제거 / 화자분리를 통한 **가용 데이터 확보 필요**

5

제언

1 배경음 제거(noise reduction)

2 화자 분리(voice separation)

5
제어

1 배경음 제거(noise reduction)

직접적으로 음성 파일이나 **numpy ndarrays**에

음향 효과를 적용해주는 **pysndfx** 라이브러리

+

음성 파일 정보를 얻어 **parameter**조정을 위해

필요한 **Librosa** 라이브러리

5
제어

1 배경음 제거(noise reduction)

pysndfx 라이브러리의 **AudioEffectsChain**함수를 이용해 6가지 방법으로 노이즈 제거한 결과 **power**를 이용한 경우가 가장 노이즈 제거 효과가 좋았음.

original.ver

r

noise

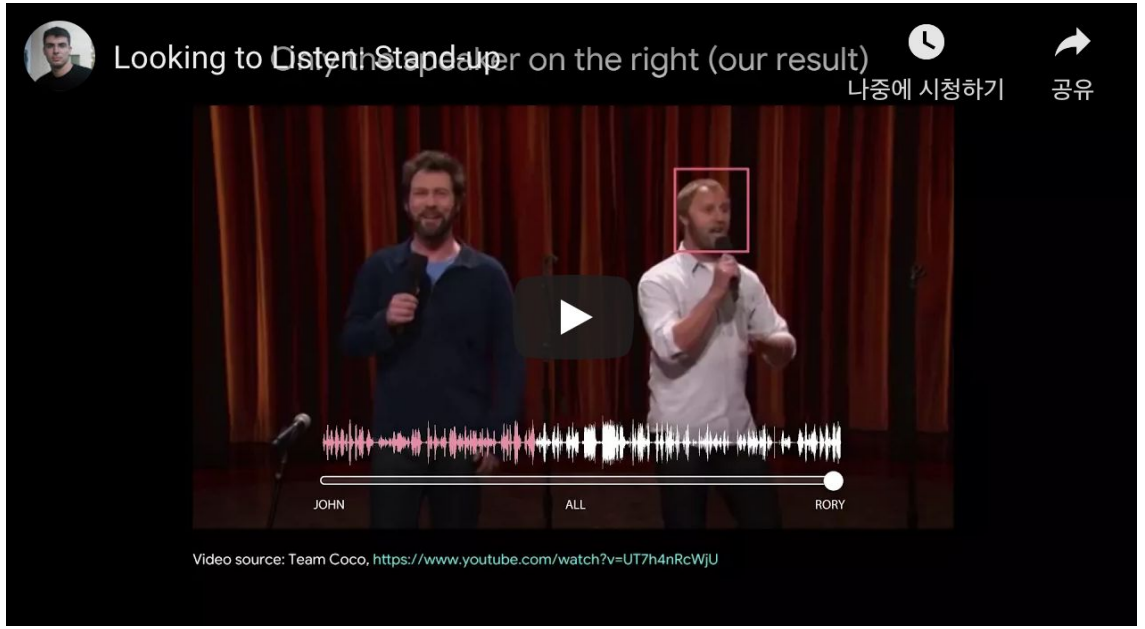
reduction.ver

5

2 화자 분리(voice separation)

Looking to Listen : Audio-Visual Speech Separation, Google Reasearch(Apr.11.2018)

- 필요한 데이터 : 화자가 등장하는 영상
- 음성 정보와 영상에 등장하는 화자의 입모양 정보을 토대로 학습
- 동영상 음성에서 하나의 음성만 분리
- 아이유 데이터 또한 영상에서 추출한 것이므로 활용가능



5
제언

- 모델 개선 측면

최종 결과물과 실제 음성을 **autoencoder 방식**을

통해 변환하는 방법 고려

THANK YOU

TOBIG`S RAHPSODY