

---

# Real English

## 영어 발음 교정 프로그램

ASAC 4기 DL 프로젝트 1조

---

01

주제 선정 배경

02

주제 소개

03

데이터 수집 및 전처리

04

모델링

05

결과

06

활용 방안

---

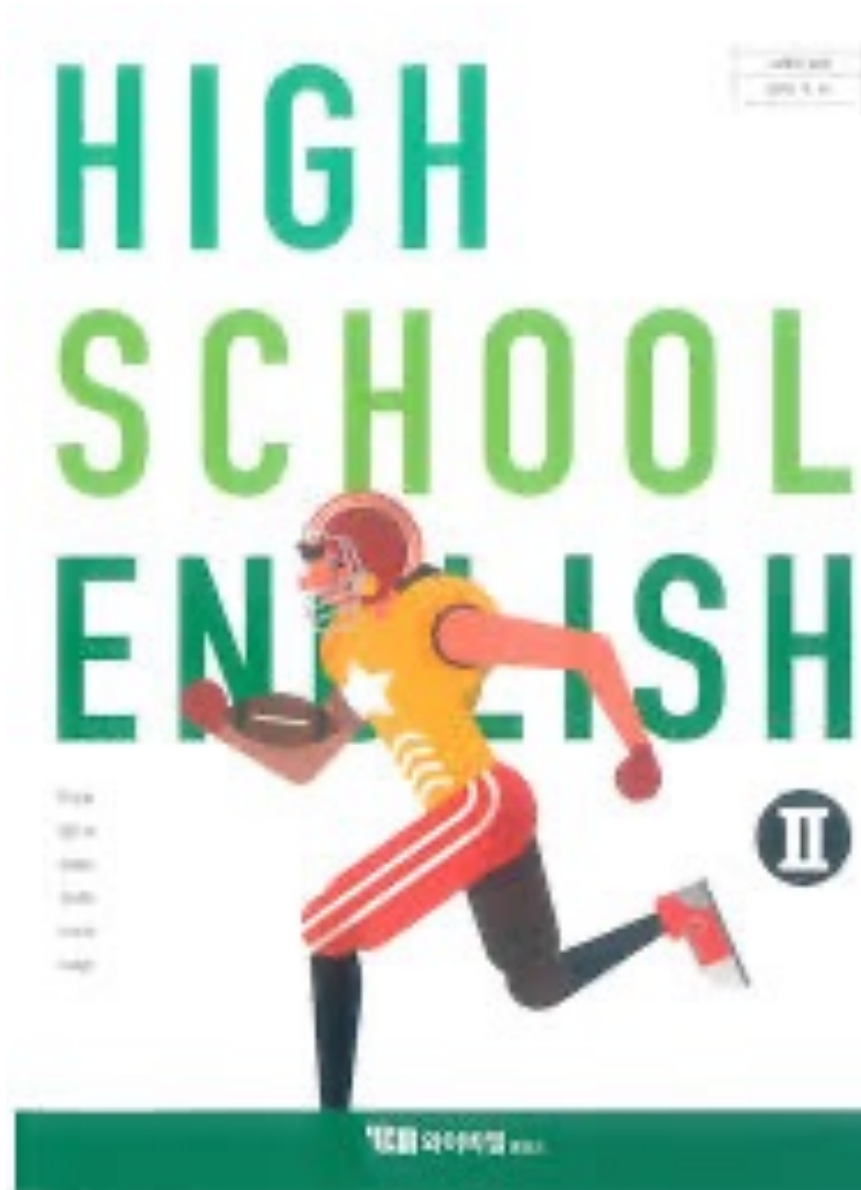
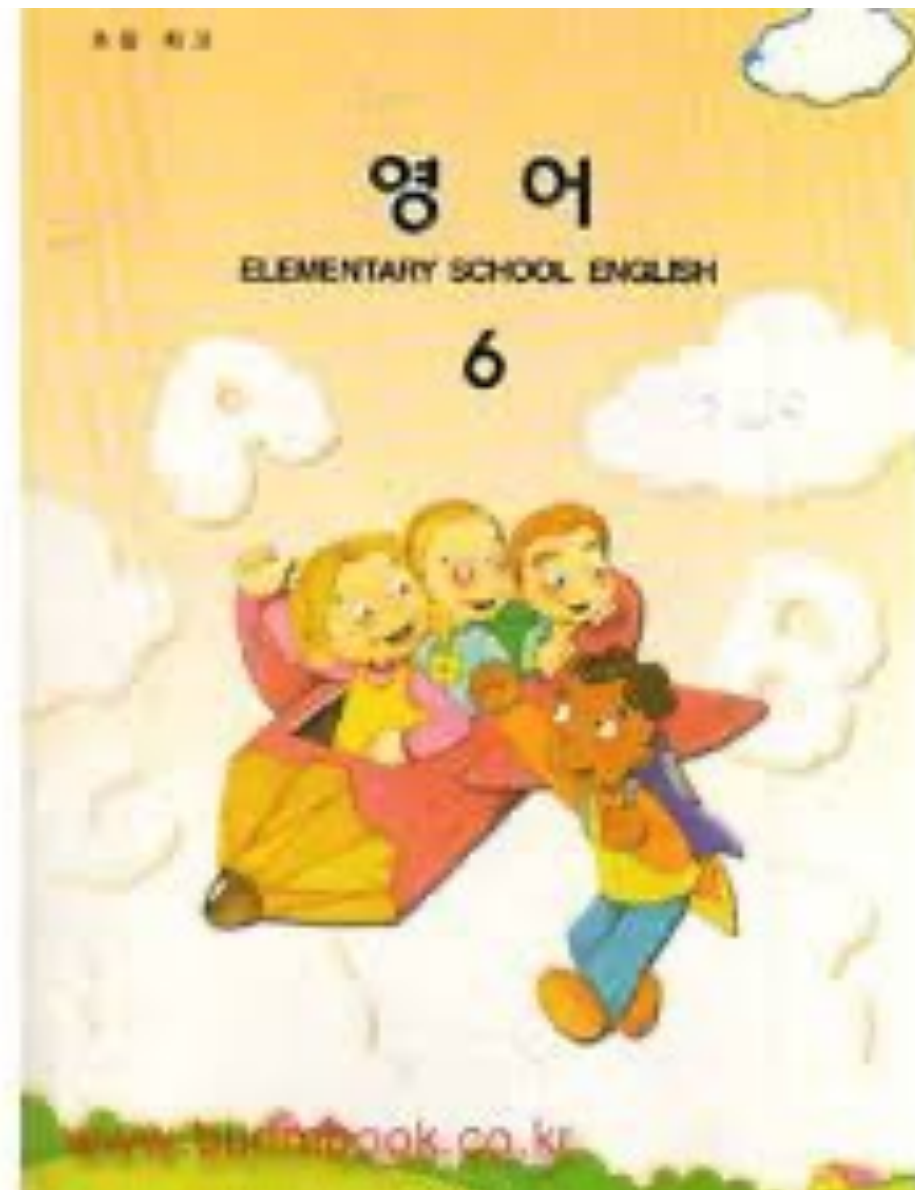
---

# 1. 주제 선정 배경

---

## 1. 주제 선정 배경

영어.. 다들 잘 하시죠..?



33. Grief is unpleasant. Would one not then be better off without it altogether? Why accept it even when the loss is real? Perhaps we should say of it what Spinoza said of regret: that whoever feels it is "twice unhappy or twice helpless." Laurence Thomas has suggested that the utility of "negative sentiments" (emotions like grief, guilt, resentment, and anger, which there is seemingly a reason to believe we might be better off without) lies in their providing a kind of guarantee of authenticity for such dispositional sentiments as love and respect. No occurrent feelings of love and respect need to be present throughout the period in which it is true that one loves or respects. One might therefore sometimes suspect, in the absence of the positive occurrent feelings, that \_\_\_\_\_ At such times, negative emotions like grief offer a kind of testimonial to the authenticity of love or respect. [3점]

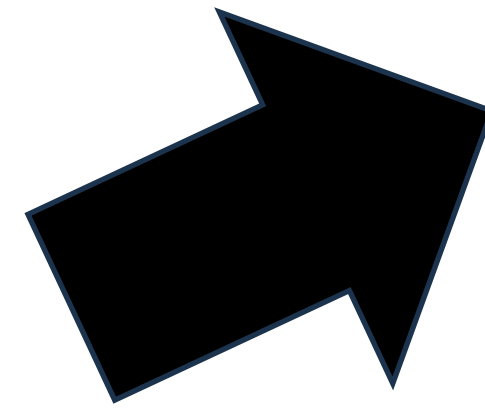
\* dispositional: 성향적인 \*\* testimonial: 증거

- ① one no longer loves
- ② one is much happier
- ③ an emotional loss can never be real
- ④ respect for oneself can be guaranteed
- ⑤ negative sentiments do not hold any longer

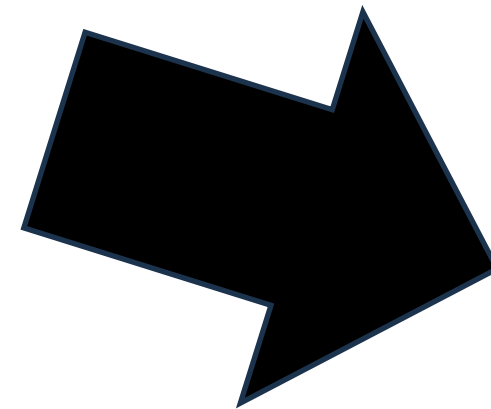
## 1. 주제 선정 배경

---

Speaking 은..?



**Konglish**  
Korean + English



**Native  
Speaking**





---

## 1. 주제 선정 배경

### 기존 영어 TTS 시스템의 한계점



‘ 이 영단어 or 영어 문장은 어떻게 읽는거지? ’  
‘ 영어 발표를 앞두고 있는데 어떻게 말해야 원어민 같지? ’

네이버	구글번역
 papago	
	



기계가 말하는 것 같다..



## 1. 주제 선정 배경

---

# So...



원어민처럼 말하고!



개선점을 찾아주는!

**영어 TTS 시스템**을 만들어보자!

---

---

## 2. 주제 소개

---



## 2. 주제 소개

---

# Real English

※ 영어 발음 교정 프로그램

### 모델링

- 원어민의 말하는 방식 학습  
(발음, 억양, 어조 등)

## Input

- Text (영어로 읽고 싶은 문장)
  - 학습자가 상기 Text 를 읽은 음성 파일
- 

## Process

- 학습시킨 모델이 아나운서 목소리로 Text 읽음
  - TTS Voice VS 학습자 Voice 음성유사도 비교
- 

## Output

- 학습시킨 모델이 Text를 읽은 오디오 파일 (원어민 아나운서st)
  - 음성 유사도 분석을 통한 개선점 제시
-

---

# 3. 데이터 수집 및 전처리

---

### 3. 데이터 수집 및 전처리

---

#### ‘까다로웠던’ 수집 데이터 조건



원어민의 말하는 방식을 학습시켜야 하기에...

1. 1명의
  2. 아나운서 직업을 가진
  3. 원어민이
  4. 20분간 발화한 음성 데이터이되
  5. 배경음 & 잡음이 없는 깔끔한 음성 데이터
-

### 3. 데이터 수집 및 전처리

---

#### 데이터 수집\_Trial 1

##### 1. YOUTUBE 해외 뉴스의 앵커 목소리 추출



- 음성 & 배경소리 분리를 통해 목소리 추출
- BUT, 완벽히 분리되지 않으며,
- 이 부분이 학습에 부정적인 영향을 끼침

 음성만 녹음된 데이터 필요성 파악!

### 3. 데이터 수집 및 전처리

---

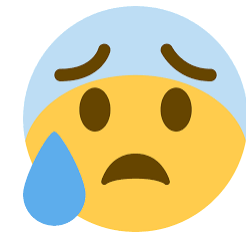
#### 데이터 수집\_Trial 2

#### 2. AI허브 영어 낭독 음성 데이터



The screenshot shows a dataset card on the AI Hub platform. On the left is a logo featuring a globe and an open book. To the right of the logo are four tags: #음성인식, #음성 합성, #기계 번역, and #LibriSpeech. Below these is a red 'NEW' badge followed by the title '다국어 통·번역 낭독체 데이터'. Under the title, it specifies '분야 한국어' and '유형 오디오'. At the bottom, it lists '구축년도: 2022', '갱신년월: 2023-12', '조회수: 2,862', '다운로드: 319', and '용량: 917.95 GB'. At the very bottom are two buttons: a red '다운로드' button and a white button with a download icon, the text '샘플 데이터', and a question mark icon.

- 1인의 깔끔한 영어 발화 음성 데이터였으나,
- 원어민이 아닌 한국인의 음성 데이터
- 본래의 취지와 맞지 않아 기각



### 3. 데이터 수집 및 전처리


---

#### 데이터 수집\_Trial 3

##### 3. 캐글 성경 낭독 음성 데이터

## The World English Bible

A large, single-speaker speech dataset in English

- 1인의 원어민 발화 음성 데이터였으나,
- 깔끔치 못한 음질과, 성경 낭독 특유의 억양으로
- 또 기각.. 

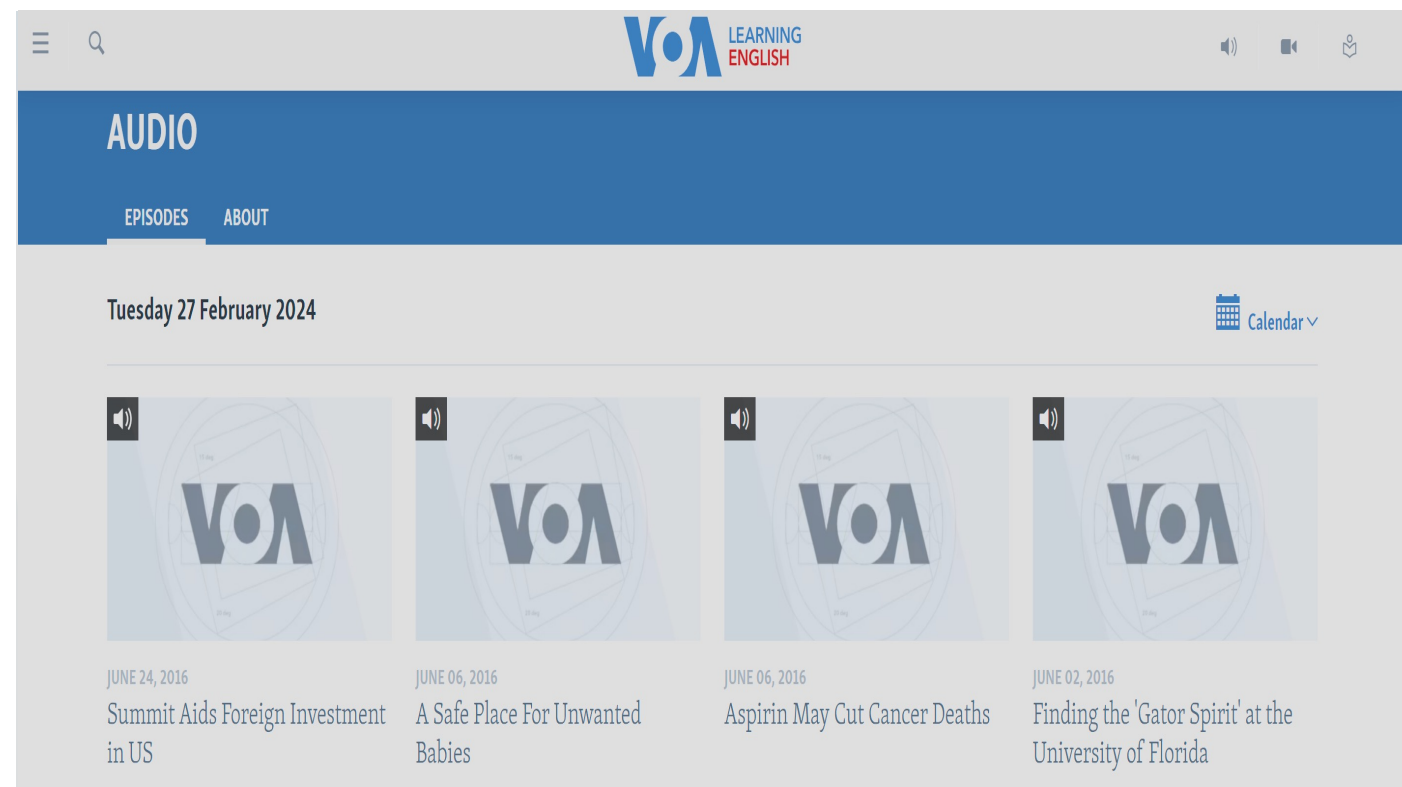


### 3. 데이터 수집 및 전처리

---

#### 데이터 수집\_Trial 4

#### 4. VOA(Voice Of America) 오디오 뉴스 데이터



\* 미국 공기업에서 운영하는 국제방송

- 1인의 깔끔한, 원어민 아나운서 음성 데이터

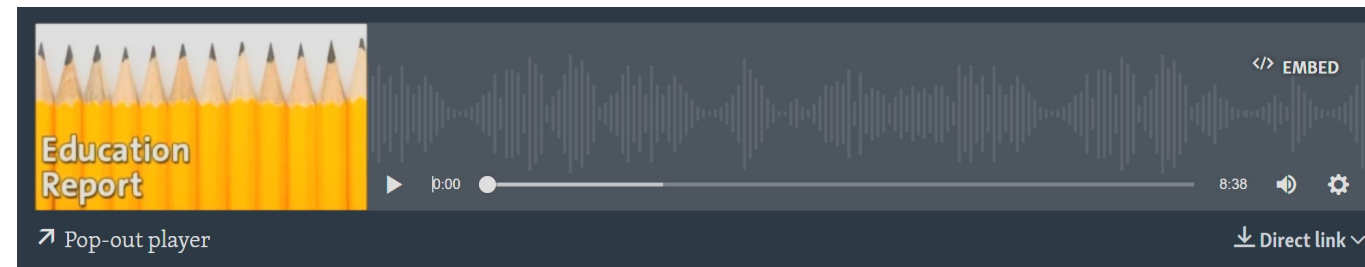
- 겿또.



### 3. 데이터 수집 및 전처리

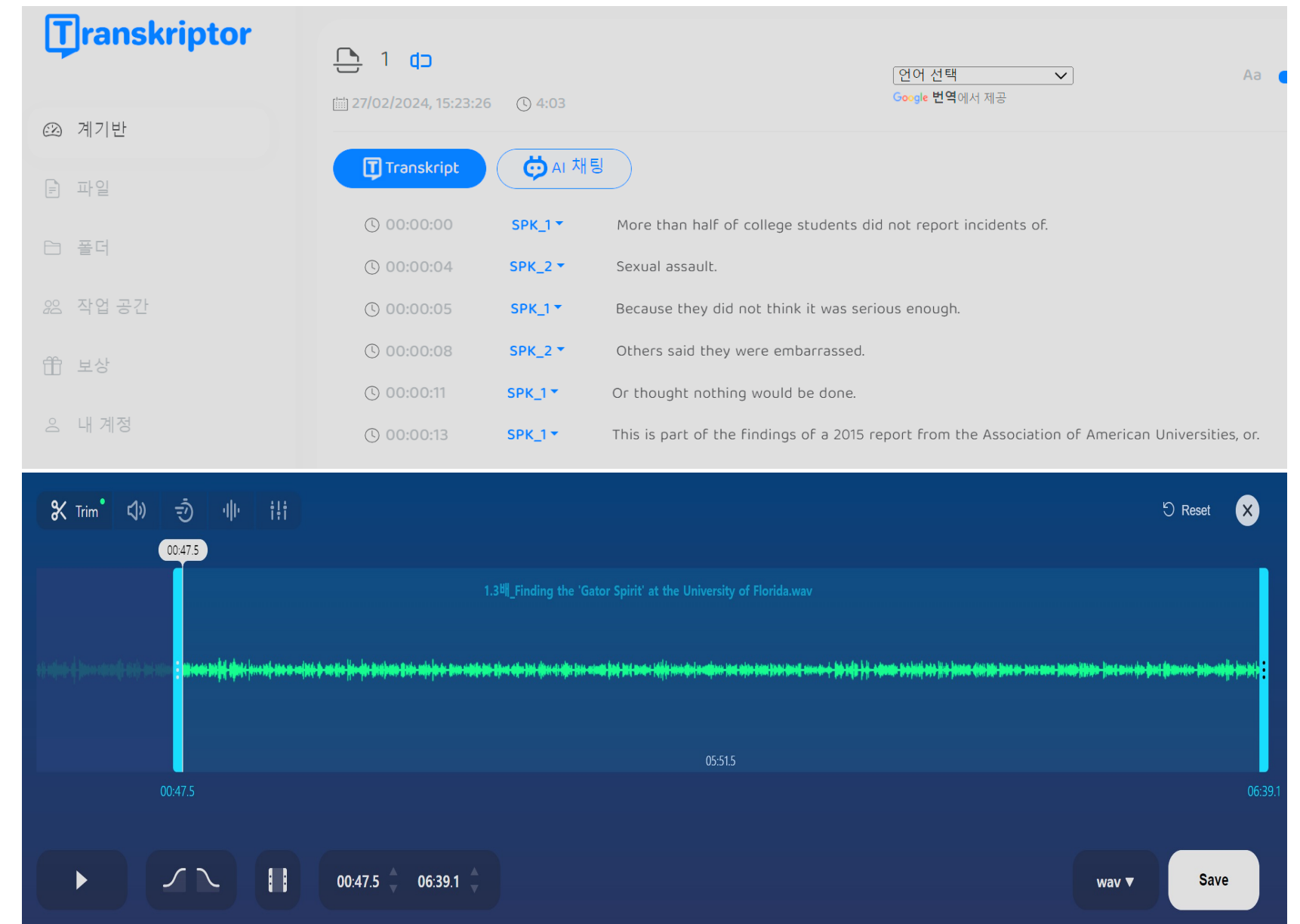
#### 데이터 전처리

## 0. Raw Data



- 5~8분 길이의 음성 데이터 3개
- mp3 -> wav 파일로 변환 진행
- 파일 크기 제약으로 인해,  
5~10초 사이의 파일로 소분 필요
- 각 파일에 대한 대본 필요

## 1. 구글 확장프로그램 사용



➡ 100개의 음성 데이터 & 대본 생성

### 3. 데이터 수집 및 전처리

---

#### 데이터 전처리

## 2. 정규화

- 숫자, %, \$ 등의 기호 영어로 변환
- 특수기호, 문장부호 삭제

## 3. Hz(헤르츠) 변경

- Sampling Rate 22050Hz로 변경
  - VITS(TTS 모델)에서 요구하는 Hz
-

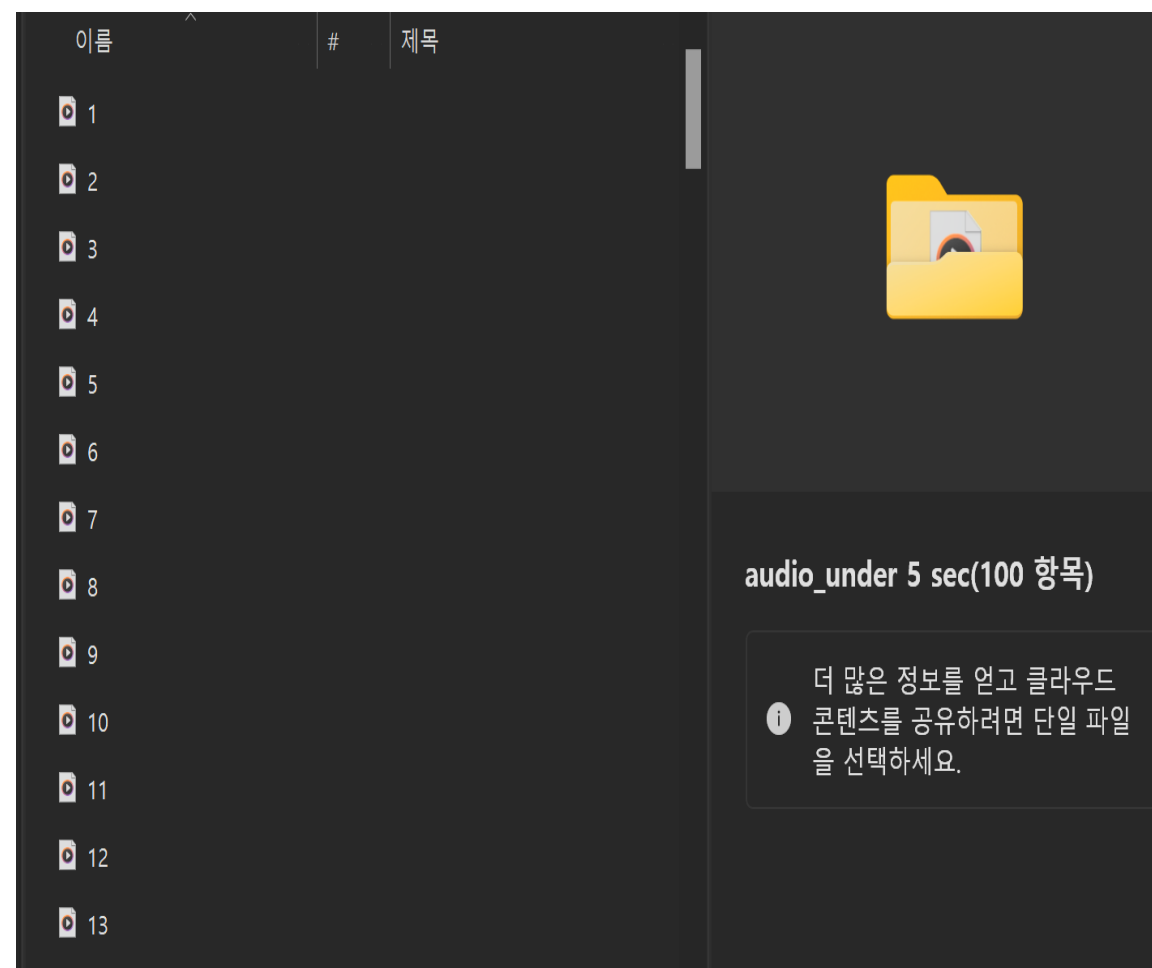
### 3. 데이터 수집 및 전처리

## 전처리 완료 데이터셋

Sample Rate: 22050 Hz  
Duration: 8.03 seconds  
Channels: Mono  
Sample Width: 16 bits

## \* 음성데이터 특징

# 100개의 음성 데이터



대본.CSV

[illegible]

---

## 4. 모델링

---

## 4. 모델링

---

### COQUI TTS 라이브러리 내 'VITS' 모델 활용



🐸 TTS is a library for advanced Text-to-Speech generation.

🚀 Pretrained models in +1100 languages.

🔧 Tools for training new models and fine-tuning existing models in any language.

📖 Utilities for dataset analysis and curation.

- Text-To-Speech 라이브러리로,  
텍스트를 음성으로 변환해주는 기능 제공

- Coqui TTS 라이브러리 내 여러 모델 중,  
적대적 학습을 사용하여,  
자연스러운 음성을 생성해주는 VITS 모델 채택

\* VITS : 카카오에서 개발한 TTS 모델

---



## 4. 모델링

### GCP 가상환경 세팅

#### VM 인스턴스

필터 속성 이름 또는 값 입력										
<input type="checkbox"/>	상태	이름 ↑	영역	권장사항	다음에서 사용 중:	내부 IP	외부 IP	연결		
<input type="checkbox"/>	✓	<a href="#">instance-20240223-sjp</a>	asia-southeast1-c			10.148.0.10 ( <a href="#">nic0</a> )	<a href="#">34.143.237.214</a> ( <a href="#">nic0</a> )	SSH	▼	⋮
<input type="checkbox"/>	✓	<a href="#">instance-20240226-1jo</a>	asia-northeast3-b			10.178.0.3 ( <a href="#">nic0</a> )	<a href="#">34.64.135.11</a> ( <a href="#">nic0</a> )	SSH	▼	⋮
<input type="checkbox"/>	✓	<a href="#">savemeplz</a>	asia-southeast1-c			10.148.0.12 ( <a href="#">nic0</a> )	<a href="#">34.126.131.96</a> ( <a href="#">nic0</a> )	SSH	▼	⋮
<input type="checkbox"/>	⊙	<a href="#">test-ubuntu</a>	us-central1-a			10.128.0.6 ( <a href="#">nic0</a> )		SSH	▼	⋮
<input type="checkbox"/>	⊙	<a href="#">tt4instance-20240222-121748</a>	asia-east2-c			10.170.0.2 ( <a href="#">nic0</a> )		SSH	▼	⋮

#### ① GCP 인스턴스 스펙

- GPU : nvidia T4, n1-standard-4
- 부팅디스크 : Ubuntu, 100GB


#### ② 환경설정

- Anaconda : Anaconda3-2022.10
- pytorch : 1.13.0
- torchvision : 0.14.0
- torchaudio : 0.13.0
- pytorch-cuda : 11.7

## 4. 모델링

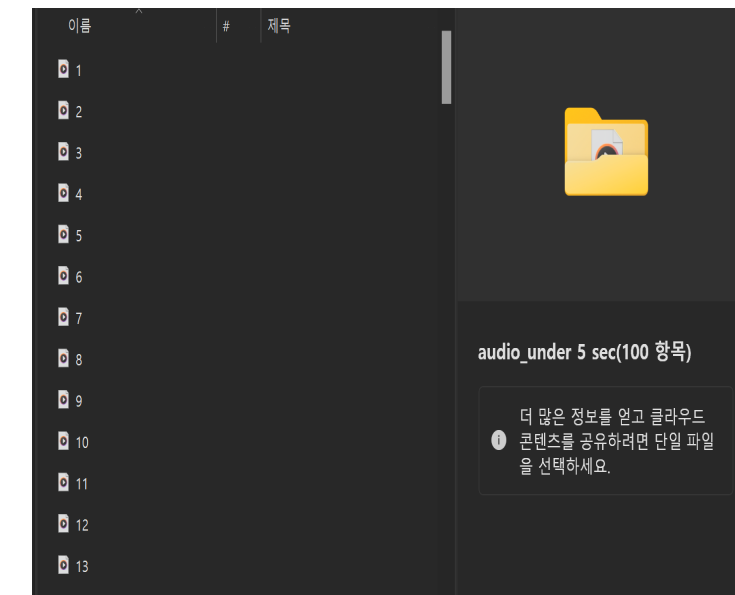


① VITS 모델 내  
사전학습 되어있는 모델 초기화



	header	header	header	header	header	header
1	B57828	2020. 4. 6.	일본프로 야구 선수	₩1,415,000	180	No.548487
2	B57828	2020. 4. 6.	일본프로 야구 선수	₩1,415,000	180	No.548487
3	B57828	2020. 4. 6.	일본프로 야구 선수	₩1,415,000	180	No.548487
4	B57828	2020. 4. 6.	일본프로 야구 선수	₩1,415,000	180	No.548487
5	B57828	2020. 4. 6.	일본프로 야구 선수	₩1,415,000	180	No.548487
6	B57828	2020. 4. 6.	일본프로 야구 선수	₩1,415,000	180	No.548487
7	B57828	2020. 4. 6.	일본프로 야구 선수	₩1,415,000	180	No.548487
8	B57828	2020. 4. 6.	일본프로 야구 선수	₩1,415,000	180	No.548487
9	B57828	2020. 4. 6.	일본프로 야구 선수	₩1,415,000	180	No.548487
10	B57828	2020. 4. 6.	일본프로 야구 선수	₩1,415,000	180	No.548487

② 아나운서 음성 데이터셋  
수집 및 전처리



③ 학습 환경 설정 및 진행

```
--> EVAL PERFORMANCE
| > avg_loader_time: 0.20614 (+0.00254)
| > avg_loss_disc: 2.17728 (-0.14163)
| > avg_loss_disc_real_0: 0.04921 (-0.08203)
| > avg_loss_disc_real_1: 0.25791 (+0.00898)
| > avg_loss_disc_real_2: 0.17484 (-0.04159)
| > avg_loss_disc_real_3: 0.18514 (-0.00242)
| > avg_loss_disc_real_4: 0.15773 (-0.14294)
| > avg_loss_disc_real_5: 0.16724 (-0.07687)
| > avg_loss_0: 2.17728 (-0.14163)
| > avg_loss_gen: 2.22632 (-0.68625)
| > avg_loss_kl: 9.13773 (+0.64543)
| > avg_loss_feat: 6.25302 (+0.12342)
| > avg_loss_mel: 28.70427 (+3.33134)
| > avg_loss_duration: 3.05029 (-0.00032)
| > avg_loss_1: 49.37163 (+3.41362)
```

④ Best Model 생성 및  
Output TTS 음성 제작

BEST  
MODEL

## 4. 모델링

### 음성 유사도 비교



TTS의 영어 발음

VS



MJ의 영어 발음

### ※ 비교 기준

1. 전달속도 - 총 발화시간(동일 Text 기준)
2. 진폭(dB) - 말의 세기(강도)
3. 억양 - 음높이(Pitch)의 변화
4. 발음 - 파이썬 STT 활용하여 정확도 확인
5. 휴지 - 발화 중 쉬는 타이밍

---

## 5. 결과

---

## 5. 결과

### 우리의 데이터셋으로 학습시킨 TTS...

※ 대본 

IB is an international organization that teaches international issues in different and more challenging ways than a traditional education does

- ➡ 부족했던 데이터셋(100)
- ➡ 부족했던 에포크 수(10,000)
- ➡ 음성 데이터와 대본의 불일치?

### Lj speech 데이터셋으로 학습시킨 TTS...

※ 대본 

IB is an international organization that teaches international issues in different and more challenging ways than a traditional education does

- ➡ 100개의 데이터
- ➡ 음성 학습을 위해 구성된 데이터셋인 만큼, 다양한 발음들이 고르게 분포되어 있어 학습에 용이
- ➡ 따라서 적은 에포크 수(5500)임에도, 높은 성능을 보여줌

## 5. 결과

### 추후 개선점...

#### ※ 공통 문제점 - 학습되지 않은 발음 기호 다수 존재

```
> Text: IB is an international organization that teaches international issues in different and more challenging ways than a traditional education does
> Text splitted to sentences.
['IB is an international organization that teaches international issues in different and more challenging ways than a traditional education does']
['<BLNK>', 'ɪ', '<BLNK>', 'b', '<BLNK>', ' ', '<BLNK>', 'ɪ', '<BLNK>', 'z', '<BLNK>', ' ', '<BLNK>', 'ə', '<BLNK>', 'n', '<BLNK>', ' ', '<BLNK>', 'ɪ', '<BLNK>', 'n', '<BLNK>', 't', '<BLNK>', 'ə', '<BLNK>', 'n', '<BLNK>', 'æ', '<BLNK>', 'f', '<BLNK>', 'ə', '<BLNK>', 'n', '<BLNK>', 'ə', '<BLNK>', 'l', '<BLNK>', ' ', '<BLNK>', 'ə', '<BLNK>', 'ɪ', '<BLNK>', 'g', '<BLNK>', 'ə', '<BLNK>', 'n', '<BLNK>', 'ə', '<BLNK>', 'z', '<BLNK>', 'e', '<BLNK>', 'ɪ', '<BLNK>', 'f', '<BLNK>', 'ə', '<BLNK>', 'n', '<BLNK>', ' ', '<BLNK>', 'ð', '<BLNK>', 'æ', '<BLNK>', 't', '<BLNK>', ' ', '<BLNK>', 't', '<BLNK>', 'i', '<BLNK>', 't', '<BLNK>', 'ʰ', '<BLNK>', 'f', '<BLNK>', 'ɪ', '<BLNK>', 'z', '<BLNK>', ' ', '<BLNK>', 'ɪ', '<BLNK>', 'n', '<BLNK>', 't', '<BLNK>', 'ə', '<BLNK>', 'n', '<BLNK>', 'æ', '<BLNK>', 'f', '<BLNK>', 'ə', '<BLNK>', 'n', '<BLNK>', 'ə', '<BLNK>', 'l', '<BLNK>', ' ', '<BLNK>', 'ɪ', '<BLNK>', 'f', '<BLNK>', 'u', '<BLNK>', 'z', '<BLNK>', ' ', '<BLNK>', 'ɪ', '<BLNK>', 'n', '<BLNK>', ' ', '<BLNK>', 'd', '<BLNK>', 'ɪ', '<BLNK>', 'f', '<BLNK>', 'ɪ', '<BLNK>', 'ə', '<BLNK>', 'n', '<BLNK>', 't', '<BLNK>', ' ', '<BLNK>', 'æ', '<BLNK>', 'n', '<BLNK>', 'd', '<BLNK>', ' ', '<BLNK>', 'm', '<BLNK>', 'ə', '<BLNK>', 'ɪ', '<BLNK>', ' ', '<BLNK>', 't', '<BLNK>', 'ʰ', '<BLNK>', 'f', '<BLNK>', 'æ', '<BLNK>', 'l', '<BLNK>', 'ɪ', '<BLNK>', 'n', '<BLNK>', 'd', '<BLNK>', 'ʰ', '<BLNK>', 'ʒ', '<BLNK>', 'ɪ', '<BLNK>', 'ŋ', '<BLNK>', ' ', '<BLNK>', 'w', '<BLNK>', 'e', '<BLNK>', 'ɪ', '<BLNK>', 'z', '<BLNK>', ' ', '<BLNK>', 'ð', '<BLNK>', 'ə', '<BLNK>', 'n', '<BLNK>', ' ', '<BLNK>', 'ə', '<BLNK>', ' ', '<BLNK>', 't', '<BLNK>', 'ɪ', '<BLNK>', 'ə', '<BLNK>', 'd', '<BLNK>', 'ɪ', '<BLNK>', 'f', '<BLNK>', 'ə', '<BLNK>', 'n', '<BLNK>', 'ə', '<BLNK>', 'l', '<BLNK>', ' ', '<BLNK>', 'ɛ', '<BLNK>', 'd', '<BLNK>', 'ʰ', '<BLNK>', 'ʒ', '<BLNK>', 'ə', '<BLNK>', 'k', '<BLNK>', 'e', '<BLNK>', 'ɪ', '<BLNK>', 'f', '<BLNK>', 'ə', '<BLNK>', 'n', '<BLNK>', ' ', '<BLNK>', 'd', '<BLNK>', 'ʌ', '<BLNK>', 'z', '<BLNK>']
[!] Character 'ʰ' not found in the vocabulary. Discarding it.
> Processing time: 6.121756076812744
> Real-time factor: 0.7363659852804018
> Saving output to tts_output.wav
```

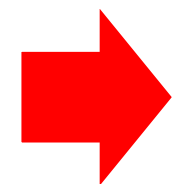
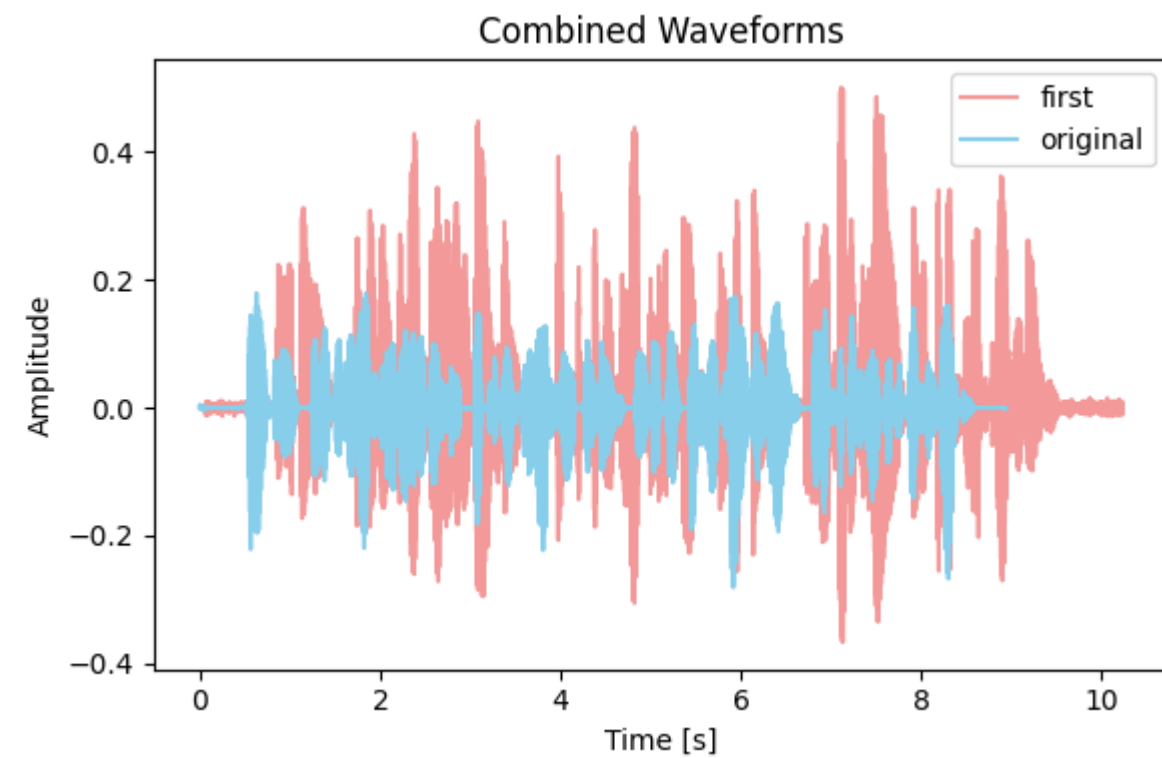
- ➡ 다양한 발음들이 고르게 분포된 충분한 데이터셋 확보
- ➡ 35,000 번 이상의 에포크 수로 학습 진행 (약 7일 ~ 10일 소요 예상)
- ➡ 이렇게 진행하게 된다면?? 🔊



## 5. 결과

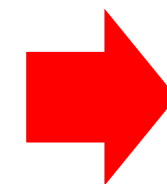
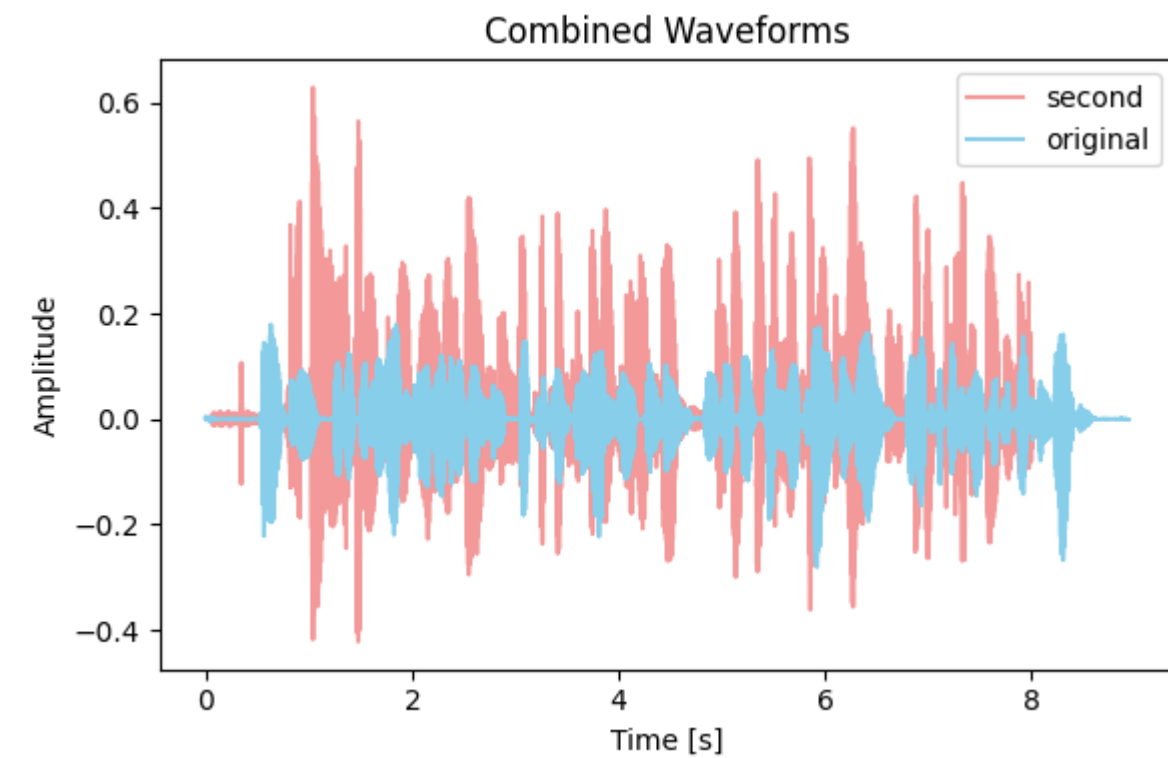
### 음성 유사도 비교

#### 1. 전달속도 - 총 발화시간(동일 Text 기준)



“ 아나운서 대비 1.3초 느리게 말하고 있습니다.  
조금만 더 빠르게 말해보세요 ”

#### 2. 진폭(dB) - 말의 세기(강도)

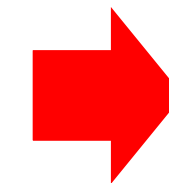
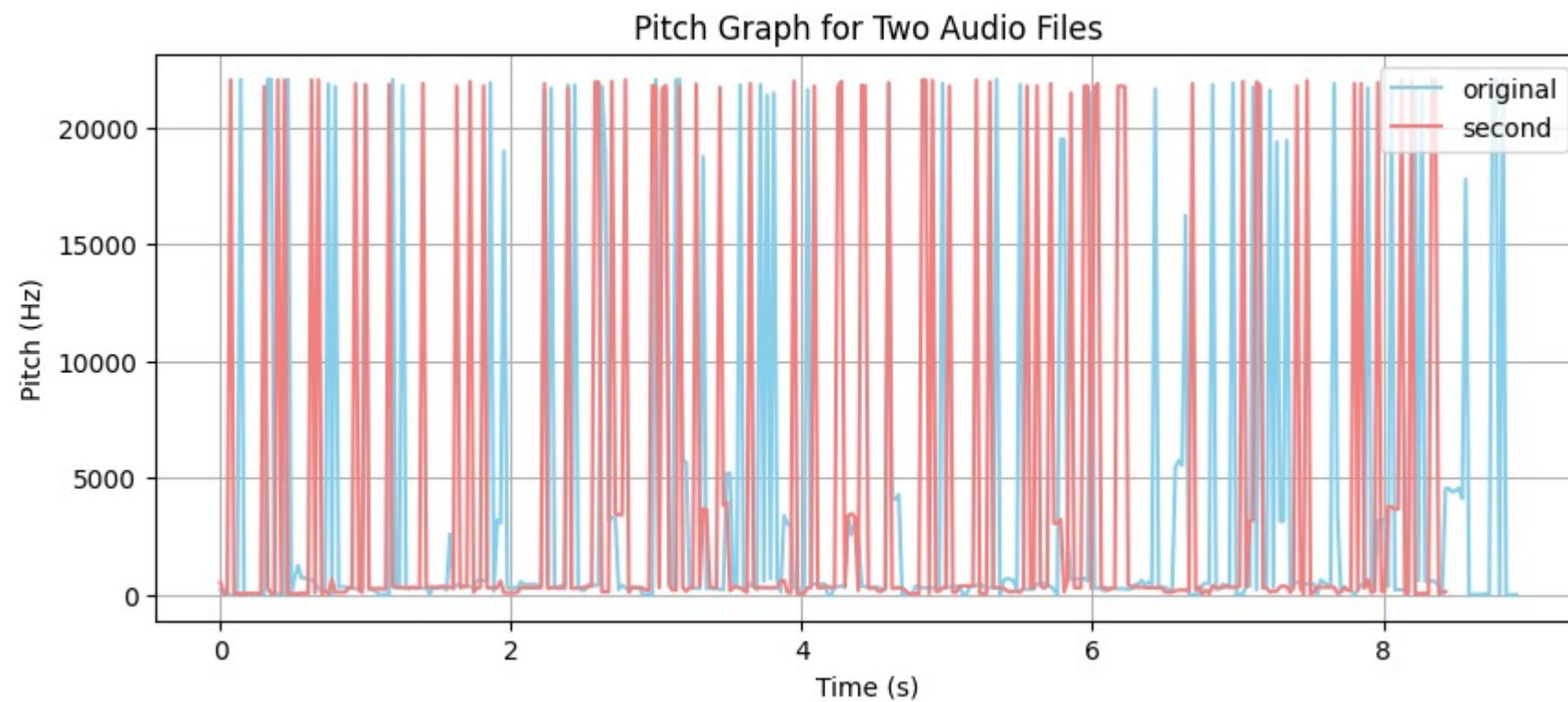


“ 아나운서 대비 너무 강하고 크게 읽고 있습니다.  
말하는 강도와 크기를 조금 줄여보세요. ”

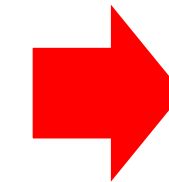
## 5. 결과

### 음성 유사도 비교

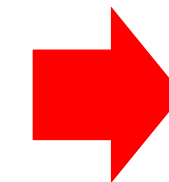
#### 3. 억양 - 음높이(Pitch)의 변화



Max 값이 같은 이유는 ,  
VITS 모델 적용을 위해 22050hz로 변환했기 때문



아나운서가 MJ보다 더 낮은 음부터 높은 음까지  
자유자재로 사용중



“ 음의 높낮이를 더욱 강조하며 읽어보세요 “

## 5. 결과

---

### 음성 유사도 비교

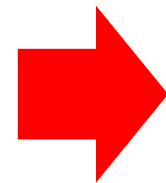
#### 4. 발음 - 파이썬 STT 활용하여 정확도 확인

##### <Original>

IB is an international organization that teaches international issues in different and more challenging ways than a traditional education does

##### <MJ>

Abby is an international organization that teaches international issues in different and more challenging ways than a traditional education does



“ IB 발음이 부정확합니다.  
원본 재청취 후 올바르게 발음해보세요 ”

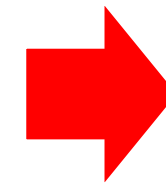
#### 5. 휴지 - 발화 중 쉬는 타이밍

##### <Original>

IB is an international ★ organization that teaches ★ international issues in different and more ★ challenging ways than a ★ traditional education does.

##### <MJ>

IB is an international ★ organization that teaches international issues in different and more challenging ways than a traditional education does.



“ 자연스러운 영어를 위해 발화 중간 중간  
추가적인 휴식이 필요합니다. 다음의 분석 내용을  
확인하시어 다시 말해보세요 ”

---

## 6. 활용방안

---

## 6. 활용방안

---

### 영어 발음 교정



#### 영어 말하기 공부

- 오픽 & 토익스피킹 준비
- 취업준비



#### 아나운서 준비

- 뉴스 앵커 / 스포츠 캐스터 등,  
원하는 업종별/방송사별로 학습 가능
- 특정 앵커 / 캐스터 / 방송인 등,  
원하는 사람의 음성 학습 가능



#### 영어 발표 준비

- 영어 면접 준비 할 때
- 예기치 못한 영어 PT 발표를 할 때
- 대본 & TTS 모델 활용하여,  
원어민처럼 말하는 법 숙지 가능

---

# 감사합니다

팀원 : 김민종, 오현진, 박소정, 한은경, 박지윤

---