



OpenAI/whisper

Gradio를 활용한 다기능 번역 모델 배포 (with whisper)

UPSTAGE AI LAB 3기

파이썬 프로젝트 16조

목차

- 01. 팀원 소개
- 02. 프로젝트 주제 소개
- 03. 이론 및 개념 설명
- 04. 프로젝트 결과 공유
- 05. 프로젝트 회고

팀원 소개



기현우

Interested in

- CV(Computer Vision)
- Medical domain AI

Role

- 프로젝트 기획 총괄 및 역할 분담
- 모델 로드 및 배포 코드 작성
- 발표 자료 제작
- 발표

Introduction

- AI의료융합전공 (4학년)
- 학부 연구원 1년 6개월 (논문)
- Dacon 각종 대회 참여

In Upstage AI Lab

- 외부 대회 의미 있는 성적 달성
- 프로젝트 경험을 쌓은 포트폴리오
- 산업기능요원 AI 직군 취업



양소현

Interested in

- NLP
- LLM
- AGI

Role

- 기획
- 실시간 자막 및 번역 관련 코드

Introduction

- 전자공학과 (4학년)
- 인턴 3회
- 프로젝트(딥러닝 진행),
창업 아이디어 등 다양한 대회
수상

In Upstage AI Lab

- AI 모델 성능 관련 대회 수상
- 코딩 테스트 마스터



이아윤

Interested in

- 미정

Role

- API 활용 크롤링
- 슬랙 봇, 카톡 메시지 전송 기능 구현

Introduction

- 체육교육 전공
- 생체역학 연구소 2년 근무

In Upstage AI Lab

- 파이썬 잘하기
- 프로젝트 경험을 쌓은 포트폴리오
- 인체 데이터 딥러닝 접목 논문 Publish



서경호

Interested in

- CV(Computer Vision)
- Medical domain AI

Role

- 모델 및 API 자료 조사
- 모델 로드 및 배포 코드 작성
- 회의록 관리와 문서 정리

Introduction

- 전기공학 졸업
- AI Reasearcher&Engineer(3년)

In Upstage AI Lab

- 대학원 진학
- AI Task Trend Follow
- AI Modeling 관련 Paper 작성

프로젝트 주제 소개

Gradio를 활용한 다기능 번역 모델 배포



Input : mic, 음원 파일, url
+ 번역 도착 언어



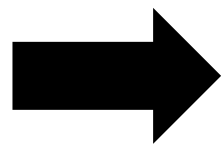
<STT + Translate>

output : detect 된 음성의 언어,
번역된 text

영상 실시간 번역

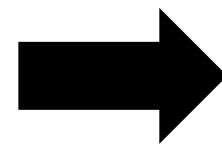


웹 브라우저에서 동영상 재생
Input : 해당 동영상의 음성
Stream으로 구현



<STT + Translate>

STT 후 Translate 진행

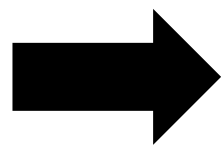


Output : Translate 된
텍스트 출력

음성 검색으로 네이버 뉴스 슬랙&카톡 메시지 전송

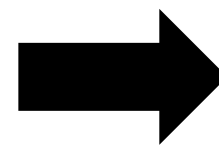


Input : mic 로 검색어 입력



<STT + Translate>

STT 후 텍스트 검색어로 전달



Output : Webdriver에서
기사 검색 후 기사 제목과 링크
Slack, 카카오톡 전송

이론 및 개념 설명

Whisper

Whisper는 Open AI에서 만들어진 자동 음성 인식(ASR)과 번역 기능을 갖춘 인공지능 모델로 다양한 언어의 음성을 텍스트로 변환하고 이를 다른 언어로 번역할 수 있다.

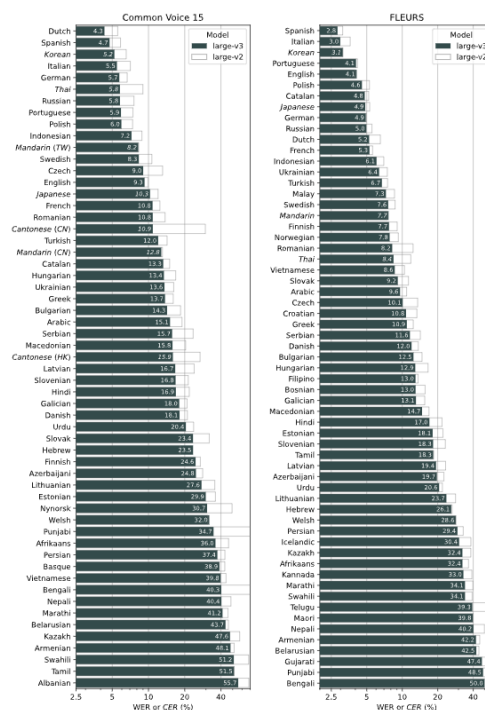
Whisper의 주요 기능

1. 음성 인식 (STT, Speech-to-text)

- 다양한 언어와 방언에서 음성을 텍스트로 변환 할 수 있다.
- 팟캐스트, 뉴스, 방송, 대화 등 다양한 유형의 오디오에서 효과적으로 작동한다.

2. 번역

- 음성을 인식한 후 그 텍스트를 다른 언어로 번역할 수 있다.
- 언어 지원 범위가 영어, 유럽어, 아시아어 등 다양한 언어를 지원한다.



Whisper가 지원하는 언어와 성능
(출처. Openai/whisper Github)

Whisper 구조 - 1

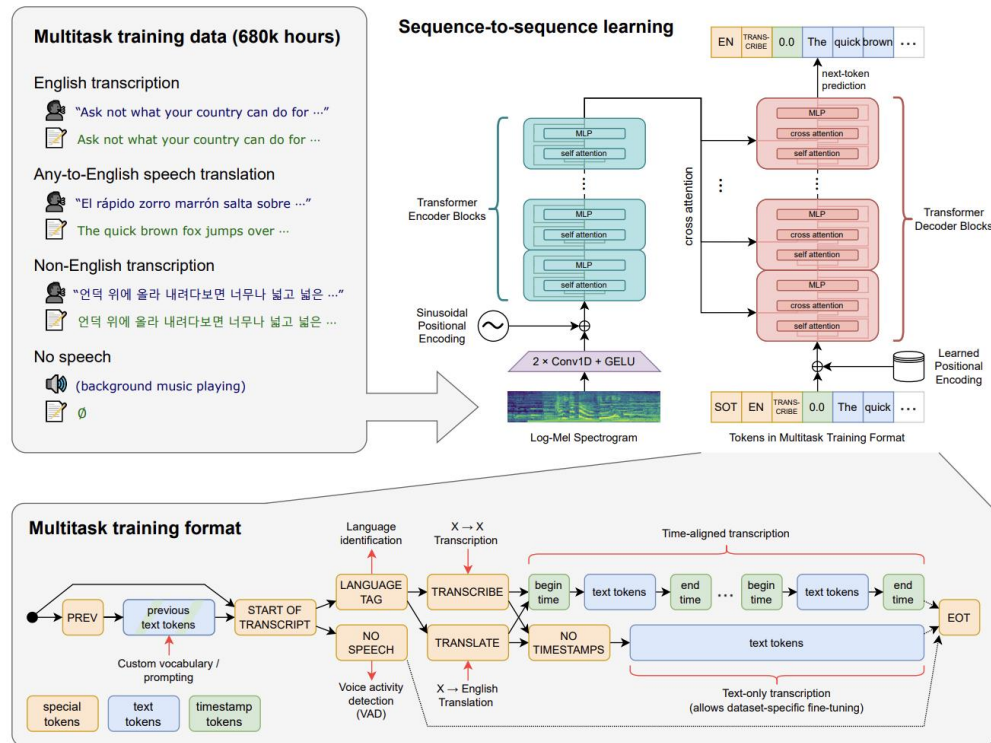


Figure 1. **Overview of our approach.** A sequence-to-sequence Transformer model is trained on many different speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. All of these tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing for a single model to replace many different stages of a traditional speech processing pipeline. The multitask training format uses a set of special tokens that serve as task specifiers or classification targets, as further explained in Section 2.3.

1. Multitask Training Data(680k hours)

Whisper는 680,000시간 분량의 다양한 음성 데이터로 구성된 매우 큰 규모의 데이터셋을 사용하여 훈련된다.

이 데이터에는 영어 transcription, 다른 언어로부터 영어로의 translation, 비영어 transcription, 음성이 없는 상황이 포함 된다.

Whisper 구조 - 2

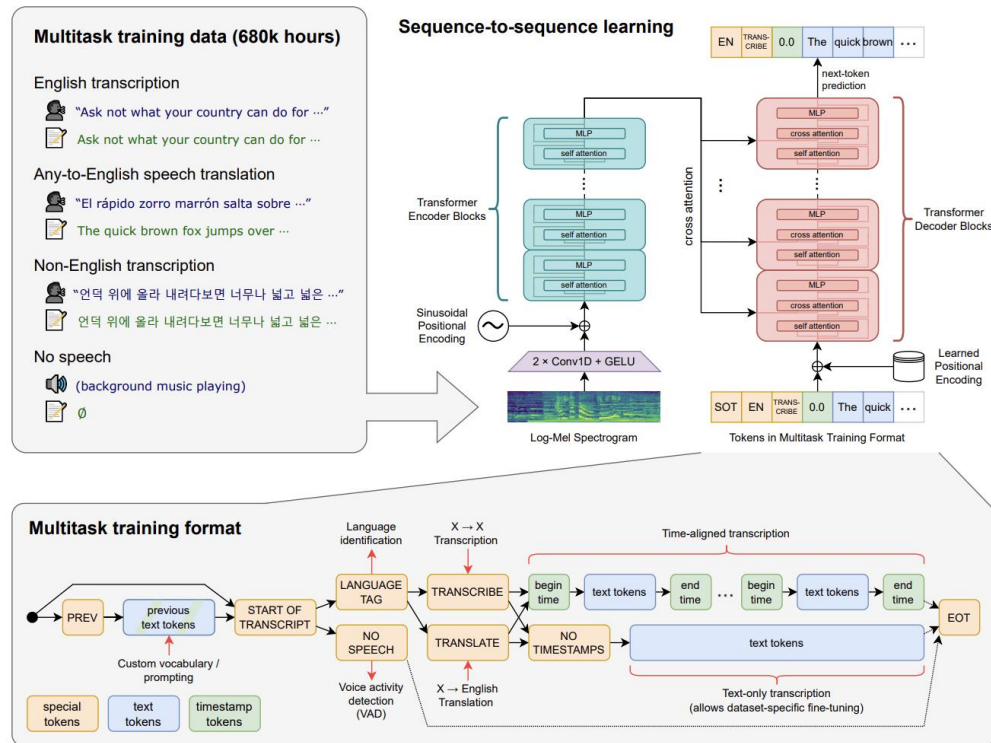


Figure 1. Overview of our approach. A sequence-to-sequence Transformer model is trained on many different speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. All of these tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing for a single model to replace many different stages of a traditional speech processing pipeline. The multitask training format uses a set of special tokens that serve as task specifiers or classification targets, as further explained in Section 2.3.

2. Sequence-to-Sequence Learning

Transformer Encoder block 과 Decoder block을 사용하여 입력으로 주어진 log-mel spectrogram에서 시작하여 최종적으로 텍스트 transcription, translation을 출력한다.

시퀀스 내의 각 단계에서 Encoder는 음성 신호를 처리하고 Decoder는 이전에 생성된 텍스트를 기반으로 다음 토큰을 예측한다.

Whisper 구조 - 3

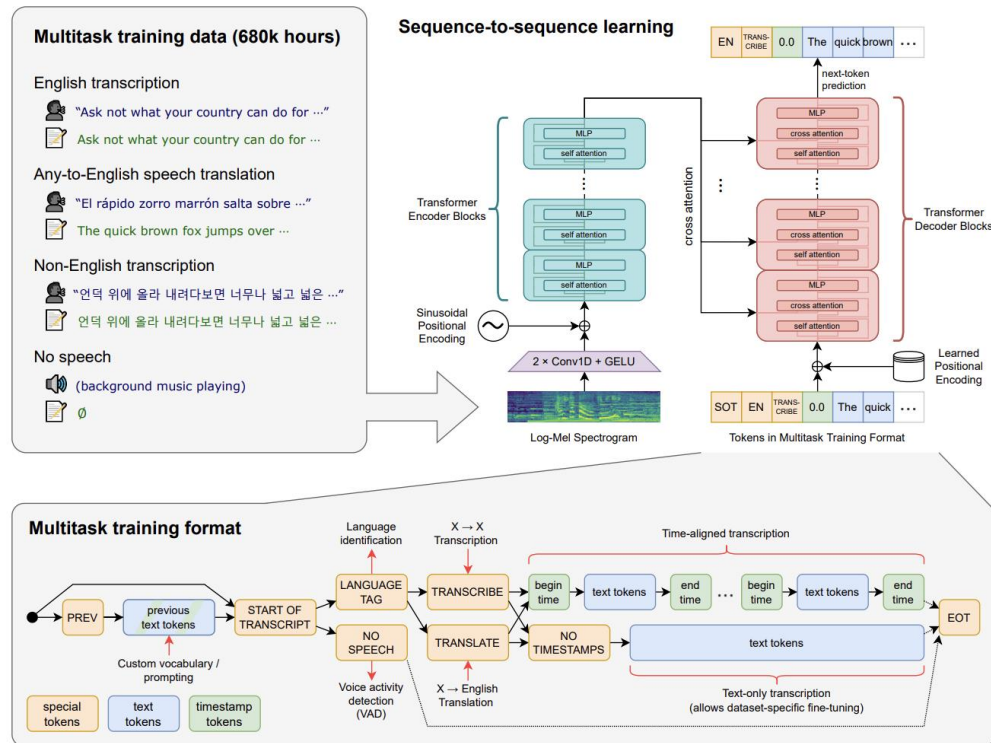


Figure 1. **Overview of our approach.** A sequence-to-sequence Transformer model is trained on many different speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. All of these tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing for a single model to replace many different stages of a traditional speech processing pipeline. The multitask training format uses a set of special tokens that serve as task specifiers or classification targets, as further explained in Section 2.3.

3. Multitask Training Format

다양한 작업을 하나의 포맷으로 통합하여 훈련하는 방식으로 언어 식별, 음성 transcription, 음성 translation이 포함된다.

각 작업은 특정 토큰으로 시작되어 해당 작업에 맞는 출력을 생성하게 된다.

(ex. "TRANSCRIBE" 태그 다음에는 transcription 수행)

Whisper 구조 - 4

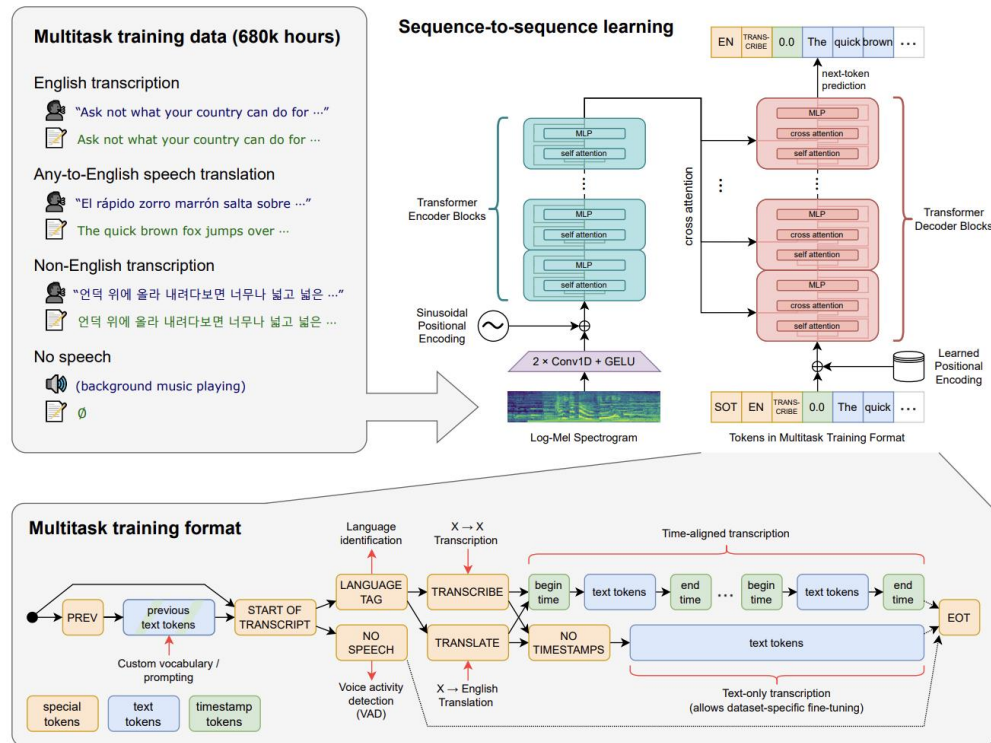


Figure 1. **Overview of our approach.** A sequence-to-sequence Transformer model is trained on many different speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. All of these tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing for a single model to replace many different stages of a traditional speech processing pipeline. The multitask training format uses a set of special tokens that serve as task specifiers or classification targets, as further explained in Section 2.3.

4. Transformer Blocks

Encoder block은 self-attention mechanism을 사용하여 입력 음성의 다양한 부분 간의 관계를 학습한다.

Decoder block은 Encoder의 출력과 이전에 생성된 텍스트 토큰을 사용하여 다음 토큰을 예측한다.

이 때 cross-attention mechanism을 사용하여 Encoder와 Decoder간의 상호 작용을 구현한다.

Whisper 구조 - 5

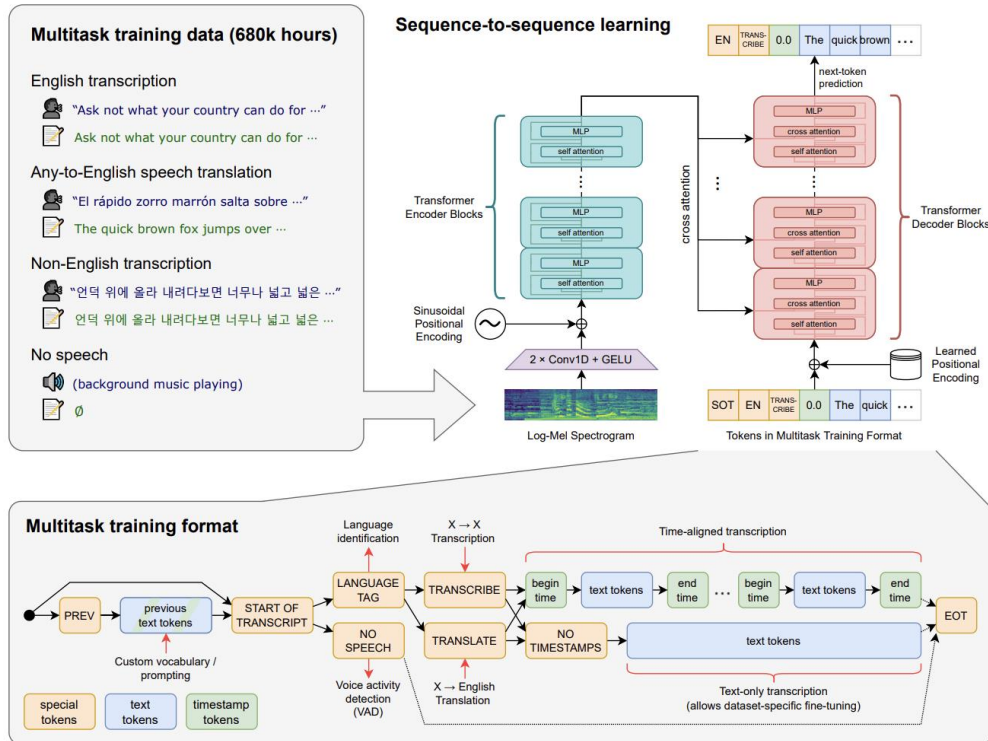


Figure 1. **Overview of our approach.** A sequence-to-sequence Transformer model is trained on many different speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. All of these tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing for a single model to replace many different stages of a traditional speech processing pipeline. The multitask training format uses a set of special tokens that serve as task specifiers or classification targets, as further explained in Section 2.3.

5. log-mel Spectrogram

음성 신호를 시각적 형태로 표현한 것으로 모델의 입력으로 사용된다.

이 Spectrogram은 CNN(Convolutional Neural Network)를 거쳐 처리된다.

Whisper 구조 - 6

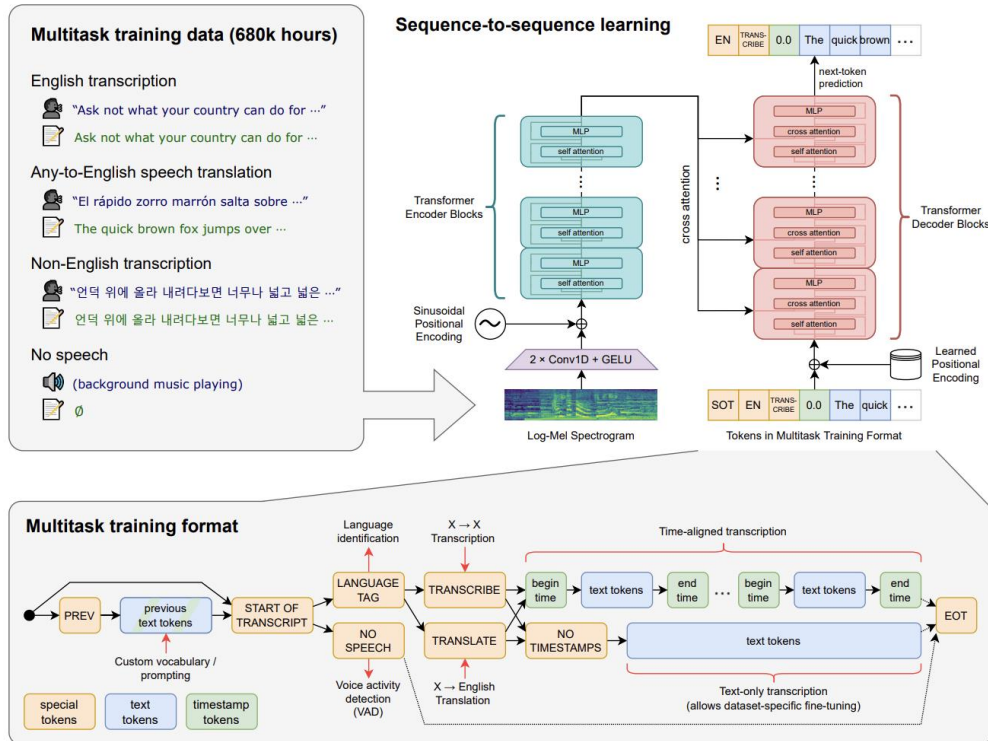


Figure 1. **Overview of our approach.** A sequence-to-sequence Transformer model is trained on many different speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. All of these tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing for a single model to replace many different stages of a traditional speech processing pipeline. The multitask training format uses a set of special tokens that serve as task specifiers or classification targets, as further explained in Section 2.3.

6. Positional Encoding

Transformer 모델에 입력 데이터의 시간적 위치 정보를 제공해 순서 정보를 이해할 수 있도록 돕는 역할을 한다.

Whisper 구조 - 7

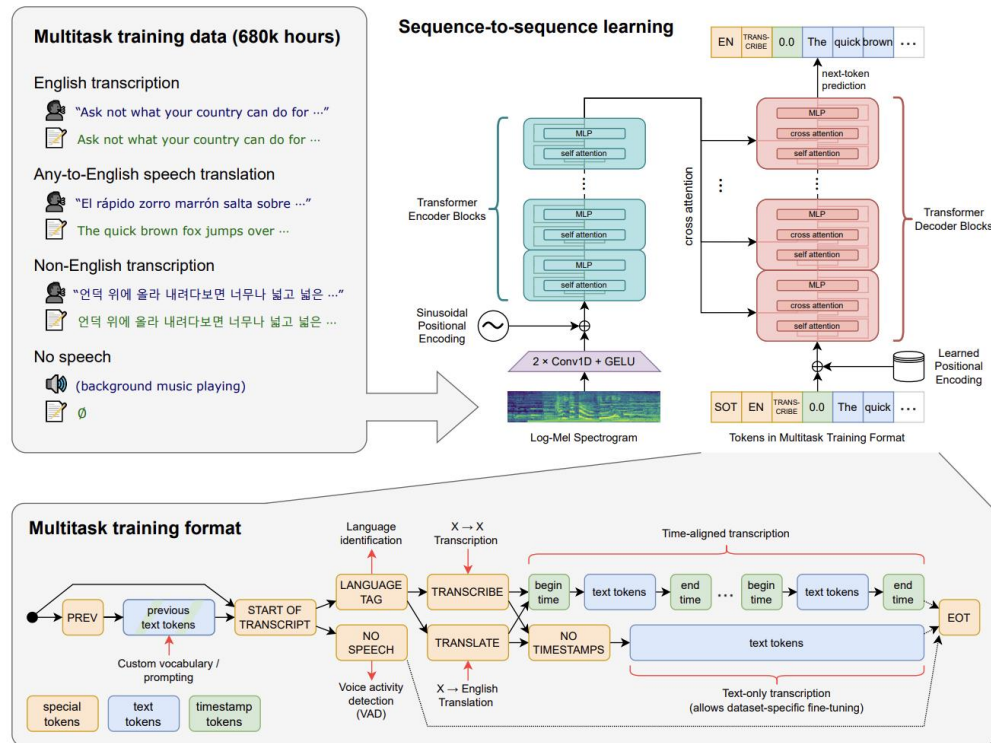


Figure 1. **Overview of our approach.** A sequence-to-sequence Transformer model is trained on many different speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. All of these tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing for a single model to replace many different stages of a traditional speech processing pipeline. The multitask training format uses a set of special tokens that serve as task specifiers or classification targets, as further explained in Section 2.3.

7. Custom Vocabulary / Special Tokens

모델은 Special Tokens를 사용하여 다양한 작업을 식별한다.

이러한 토큰들은 특정 작업을 위한 지시어 역할을 하며 모델이 어떤 작업을 수행해야 하는지 알려준다.

(ex. <|startoftranscript|> : 이 토큰은 transcription의 작업 시작을 나타낸다.)

Whisper 구조 - 8

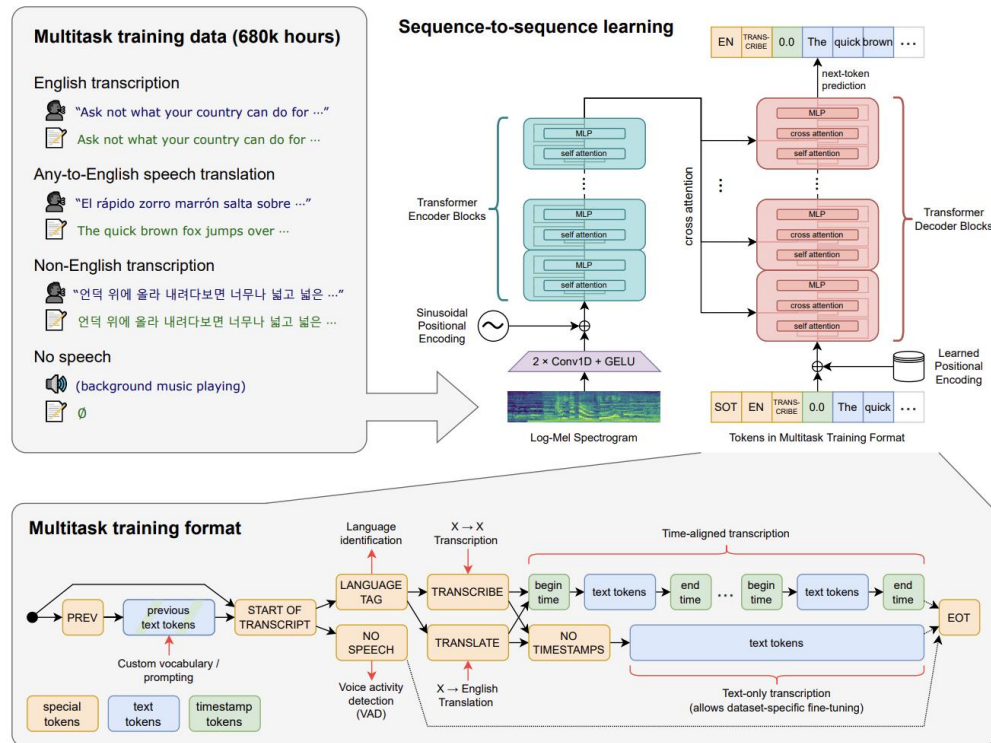


Figure 1. **Overview of our approach.** A sequence-to-sequence Transformer model is trained on many different speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. All of these tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing for a single model to replace many different stages of a traditional speech processing pipeline. The multitask training format uses a set of special tokens that serve as task specifiers or classification targets, as further explained in Section 2.3.

8. Training Format and Tokens

모델 훈련 시 다양한 형식의 데이터를 처리하기 위해 특정한 포맷이 사용된다.

이 포맷은 언어 식별, transcription, translate 등 다양한 작업을 포함하며 각 작업을 나타내기 위해 특정한 토큰이 사용되어 데이터셋을 통해 모델에게 제공된다.

Whisper 구조 - 9

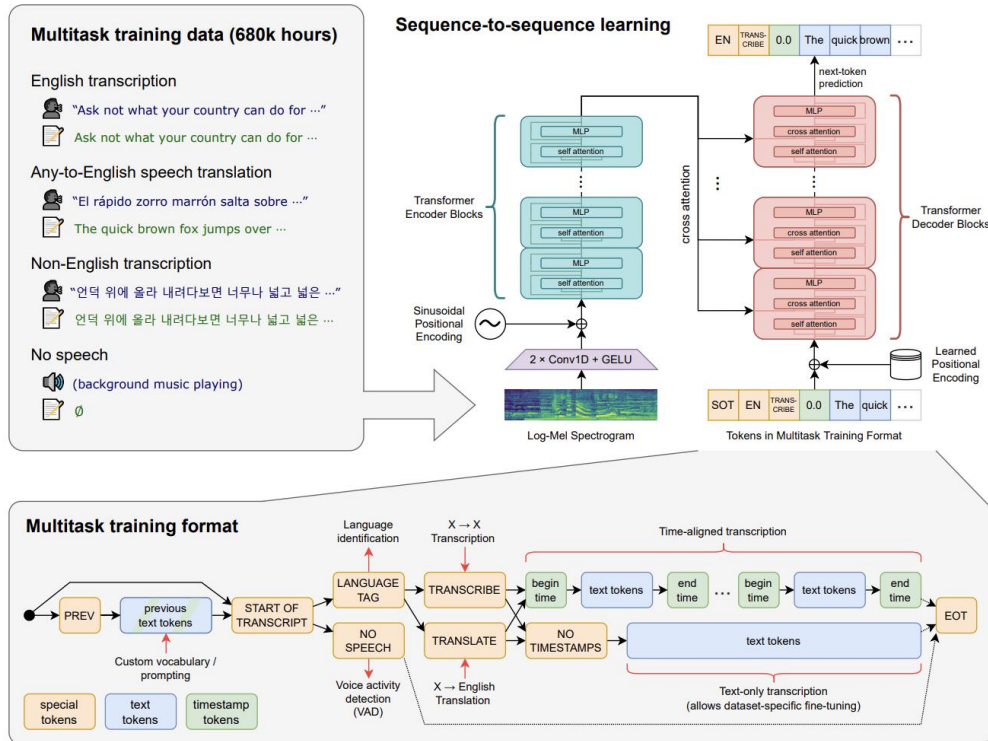


Figure 1. **Overview of our approach.** A sequence-to-sequence Transformer model is trained on many different speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. All of these tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing for a single model to replace many different stages of a traditional speech processing pipeline. The multitask training format uses a set of special tokens that serve as task specifiers or classification targets, as further explained in Section 2.3.

9. Language Identification

음성 신호에서 언어를 식별하는 과정으로 음성 번역 작업에 앞서 모델이 어떤 언어로 번역을 시작해야 하는지를 결정한다.

Whisper 구조 - 10

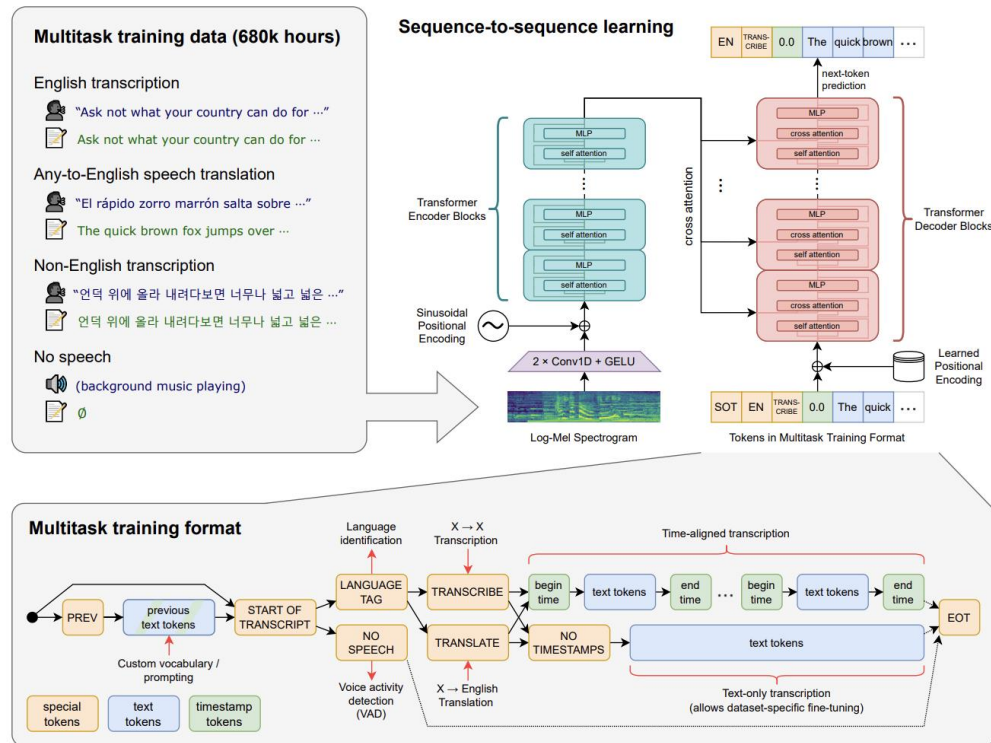


Figure 1. **Overview of our approach.** A sequence-to-sequence Transformer model is trained on many different speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. All of these tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing for a single model to replace many different stages of a traditional speech processing pipeline. The multitask training format uses a set of special tokens that serve as task specifiers or classification targets, as further explained in Section 2.3.

10. Voice Activity Detection(VAD)

오디오에서 음성이 시작하고 끝나는 부분을 식별하는 과정으로 무음 구간을 구별하고 필요한 경우에만 모델이 음성 처리를 수행하도록 한다.

Gradio

Gradio는 머신러닝 모델을 쉽게 공유하고 시연할 수 있게 해주는 오픈소스 라이브러리로 복잡한 설정 과정 없이 머신러닝 모델을 웹 인터페이스로 빠르게 전환할 수 있다.

Gradio의 주요 특징

1. 간편한 인터페이스 생성

- Gradio는 텍스트, 이미지, 오디오, 비디오 등 다양한 입력 및 출력 형식을 지원한다.

2. 통합과 호환성

- TensorFlow, Pytorch 등 주요 머신러닝 프레임워크와 호환되어 모델 개발자가 이미 사용하고 있는 도구들과 쉽게 연동할 수 있다.

3. 공유와 배포
















- 생성된 인터페이스는 URL을 통해 공유될 수 있으며 이 URL을 통해 다른 사용자들이 웹 브라우저에서 모델을 직접 사용해볼 수 있다.

프로젝트 결과 공유

1. Whisper(STT) + Marian MT(Translate)

 bob80333/speechbrain_ja2en_st_63M_yt600h
 Automatic Speech Recognition • Updated Jan 14, 2022 •  8 •  1
 espnet/brianyan918_iwslt22_dialect_st_transformer_fisherlike_4gpu_bbins16m_f...
Updated Feb 9, 2022 •  5
 espnet/brianyan918_iwslt22_dialect_train_st_conformer_ctc0.3_lr2e-3_warmup15...
Updated Feb 9, 2022 •  5
 facebook/s2t-medium-mustc-multilingual-st
 Automatic Speech Recognition • Updated Jan 25, 2023 •  931 •  3
 facebook/s2t-small-covost2-ca-en-st
 Automatic Speech Recognition • Updated Jan 25, 2023 •  27

Whisper의 경우 Translate보다 Transcription에 더 특화된 모델이라 다른 Speech translator를 보면 Whisper로 STT를 수행한 후 출력된 텍스트를 Translator모델에 입력해 번역이 되게 하였다.

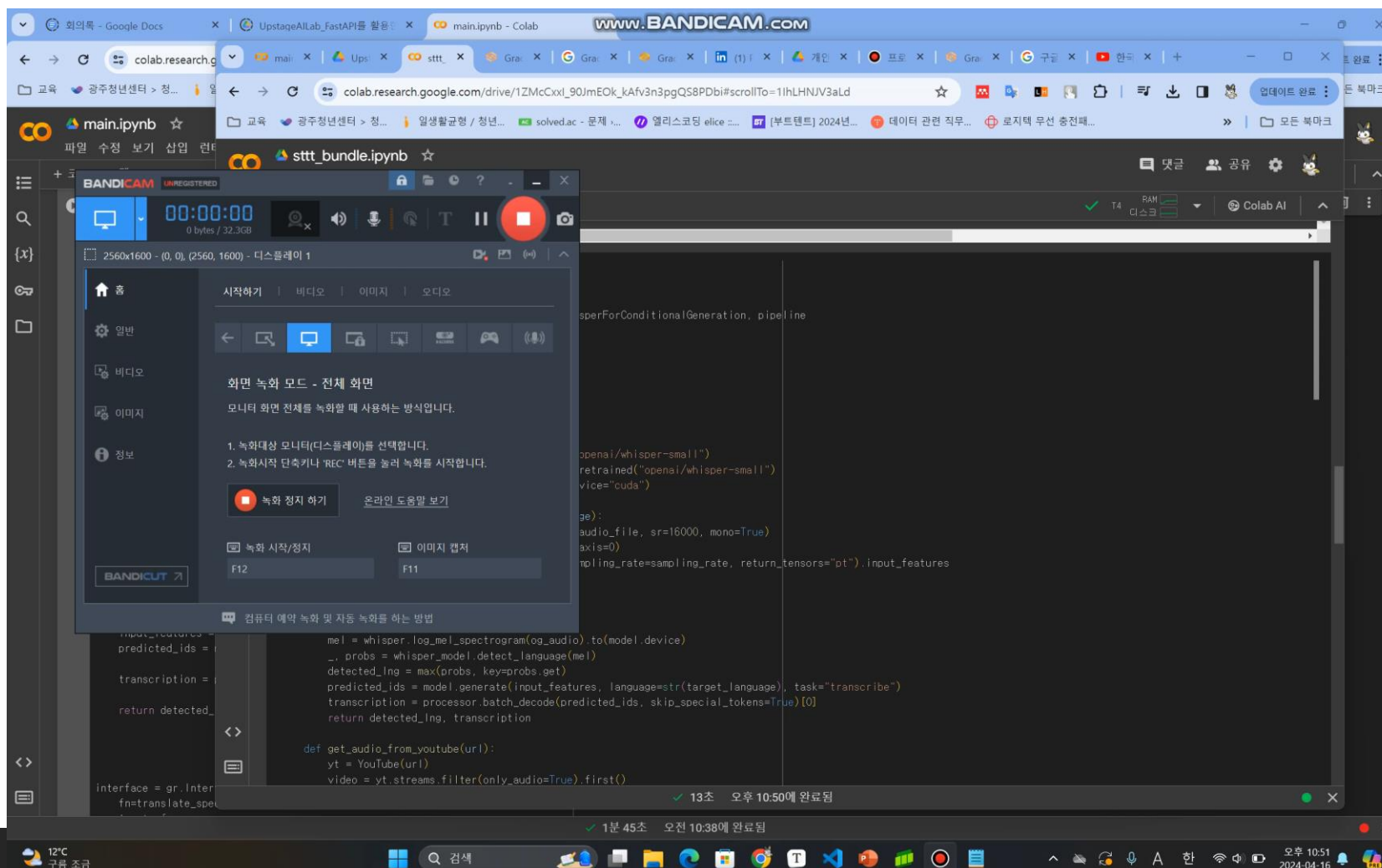
 Models 1442 	↑↓ Sort: Recently updated
 Helsinki-NLP/fin-simple-mBART	Updated 10 days ago
 Helsinki-NLP/simple-finnish-gpt3-xl	 Text Generation • Updated 10 days ago •  17 •  1
 Helsinki-NLP/opus-mt-en-fr	 Translation • Updated Feb 15 •  96.4k •  31
 Helsinki-NLP/opus-mt-sv-en	 Translation • Updated Feb 15 •  1.23M •  9

BartForConditionalGeneration 기반의 Marian MT 모델이 지원하는 언어 경로가 1442가지라 해당 모델로 Translate를 시도했으나 pretrained된 언어 경로 중 우리가 원했던 한, 중, 일 경로에 공백이 생겨 Translator를 Whisper로 수정하여 STT, Translate 모두 Whisper로 진행하였습니다.

1. Whisper(STT + Translate)

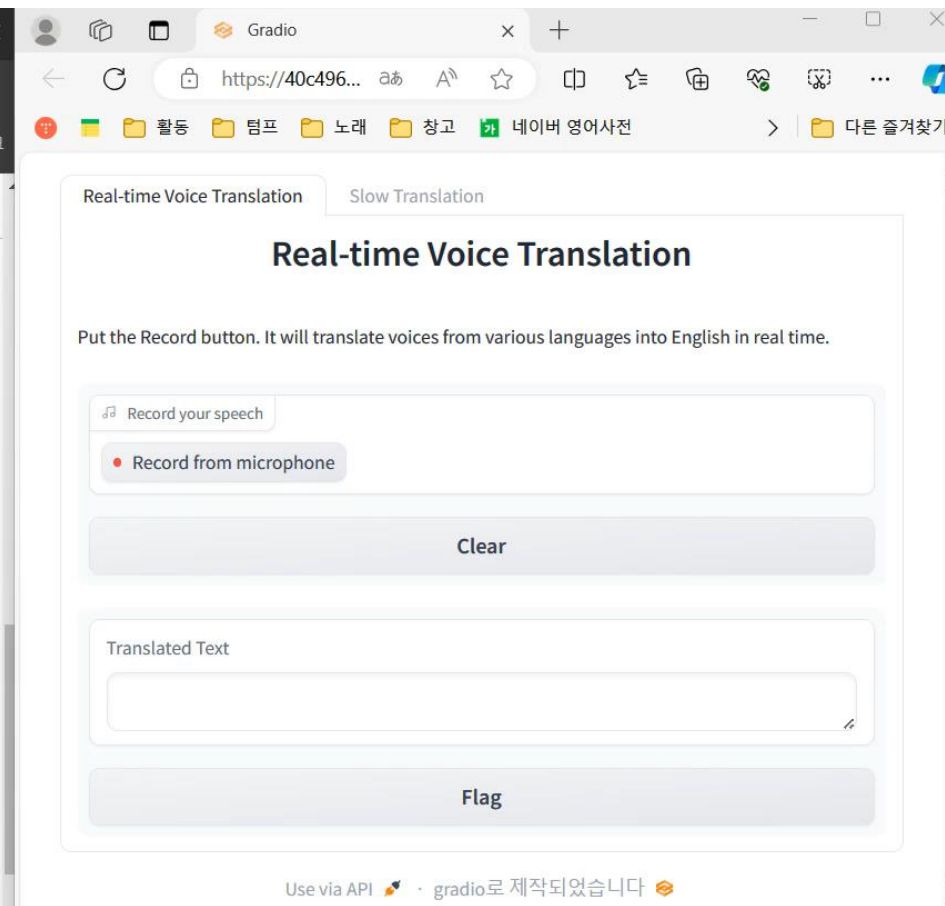
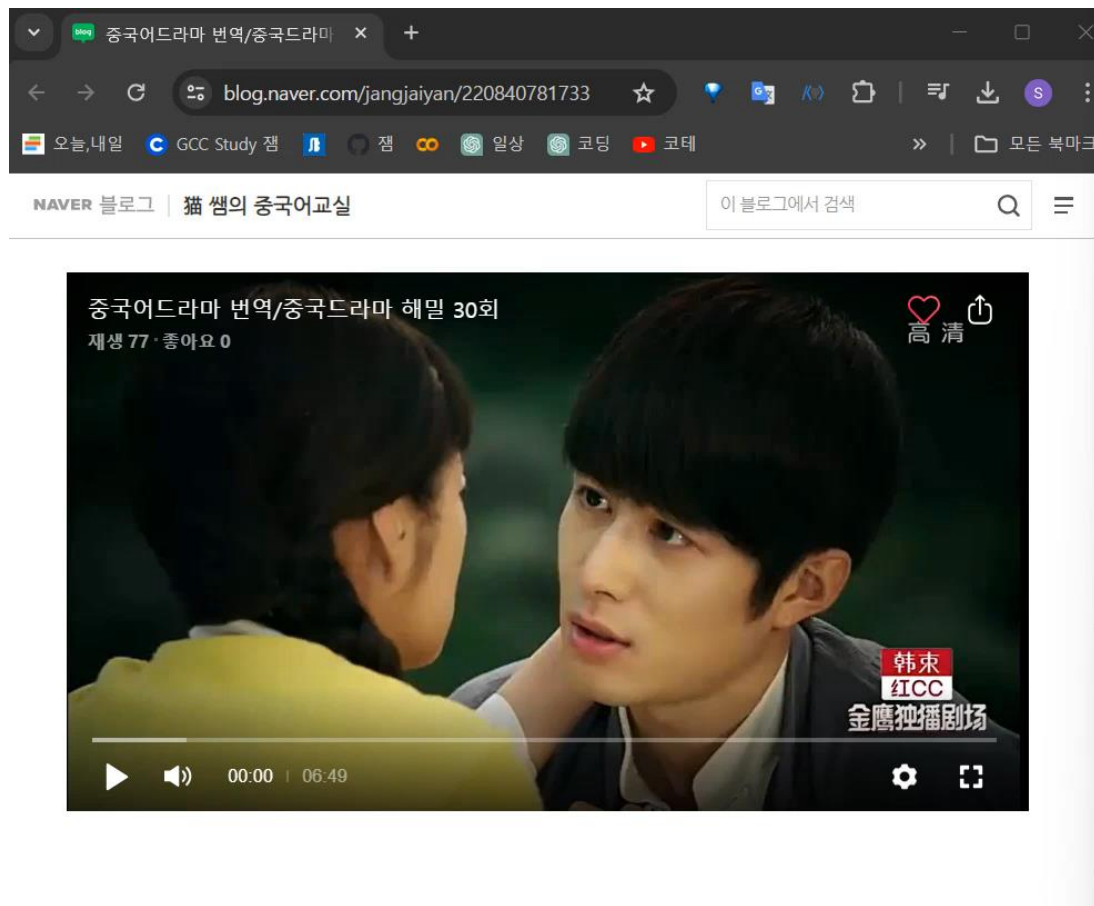
첫 기획은 '마이크로 입력되는 음성을 번역한다.' 였으나 프로젝트 진행 2일차에 gradio로 웹에 모델을 로드하는 것 까지 완료되어 기능을 여러가지 추가해 '다기능 번역 모델을 만들자'로 변경되어 음성 파일 업로드 후 번역과 유튜브 URL로 해당 영상 번역까지 추가하게 되었습니다.

1. 다기능 번역 모델 demo 영상



2. 영상 실시간 번역 demo 영상

영상 출처 : <https://blog.naver.com/jangjaiyan/220840781733>



2. 영상 실시간 번역 demo 영상

영상 출처 : <https://www.facebook.com/kjfestivalseoul/videos/882214861851606/>

The image shows two side-by-side browser windows. The left window displays a Facebook video player with a video of a man in a suit. The right window shows a Gradio web interface for real-time voice translation. The interface includes tabs for 'Real-time Voice Translation' and 'Slow Translation', a 'Record your speech' button, a 'Record from microphone' button, a 'Clear' button, a 'Translated Text' input field, and a 'Flag' button. At the bottom, it says 'Use via API' and 'gradio로 제작되었습니다'.

facebook.com/kjfestivalseoul/videos/882214861851606/

facebook

로그인

동영상 홈 라이브 릴스 프로그램 ...

동영상 검색

Real-time Voice Translation Slow Translation

Real-time Voice Translation

Put the Record button. It will translate voices from various languages into English in real time.

Record your speech

Record from microphone

Clear

Translated Text

Flag

Use via API · gradio로 제작되었습니다

2. 영상 실시간 번역 demo 영상

The image displays two side-by-side browser windows demonstrating real-time voice translation.

Left Window (Google Translate):

- Search bar: 번역기 - Google 검색
- URL: google.com/search?q=번역기&tscv=ce68b...
- Language selection: 한국어 (Korean) to 독일어 (German)
- Input text: 안녕하세요 저는 양 소현이라고 합니다. 당신의 이름은 무엇 입니까?
- Output text: Hallo, mein Name ist Sohyun Yang. Wie heißt du?
- Search results: 검색결과 약 9,700,000개 (0.21초)

Right Window (Gradio):

- Tab: Real-time Voice Translation
- Instruction: Put the Record button. It will translate voices from various languages into English in real time.
- Buttons: Record your speech, Record from microphone, Clear
- Output field: Translated Text
- Buttons: Flag
- Footer: Use via API · gradio로 제작되었습니다

2. 영상 실시간 번역 demo 영상

Real-time Voice Translation

Slow Translation

Real-time Voice Translation

Put the Record button. It will translate voices from various languages into English in real time.

Record your speech

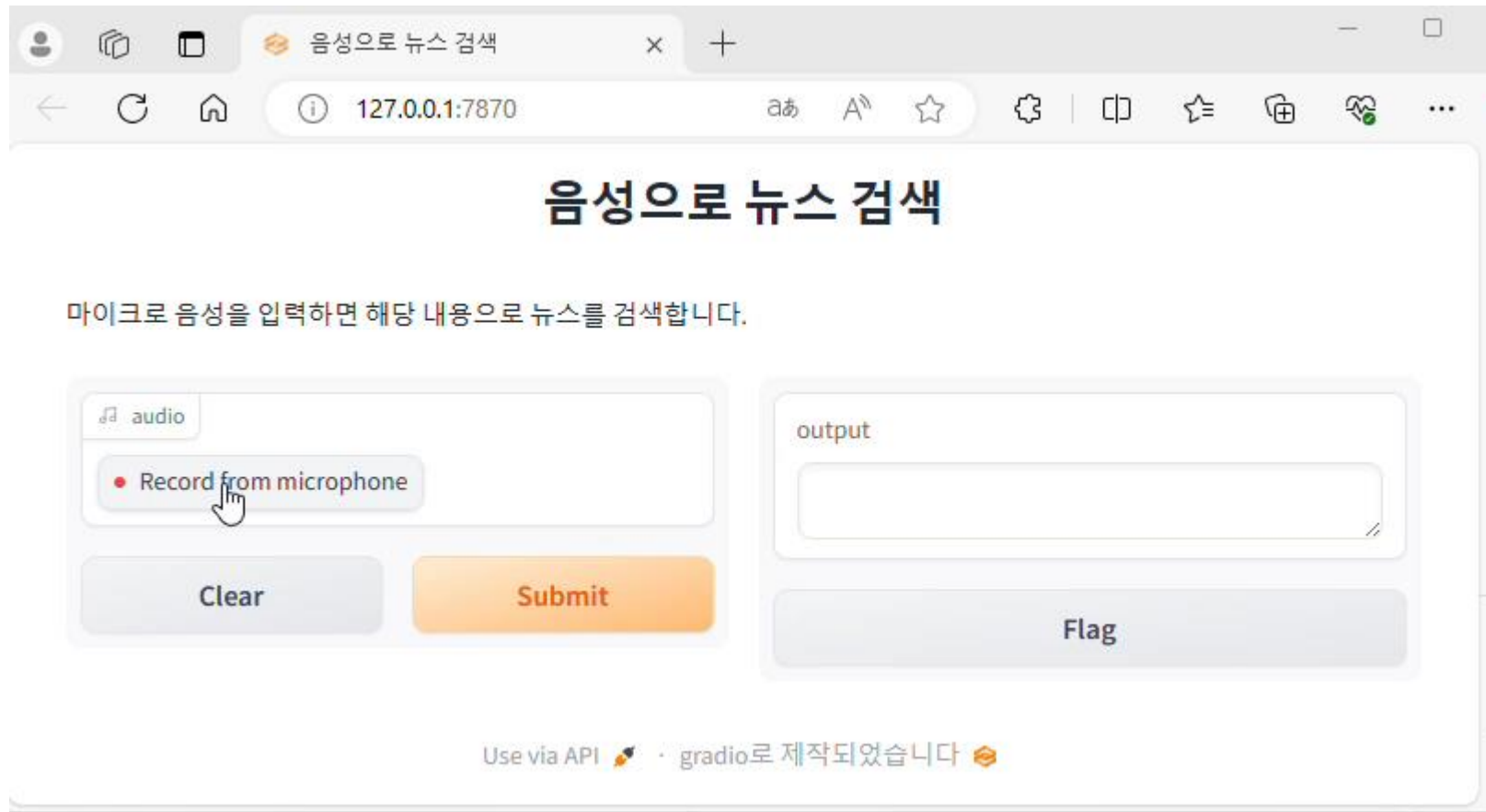
Record from microphone

Clear

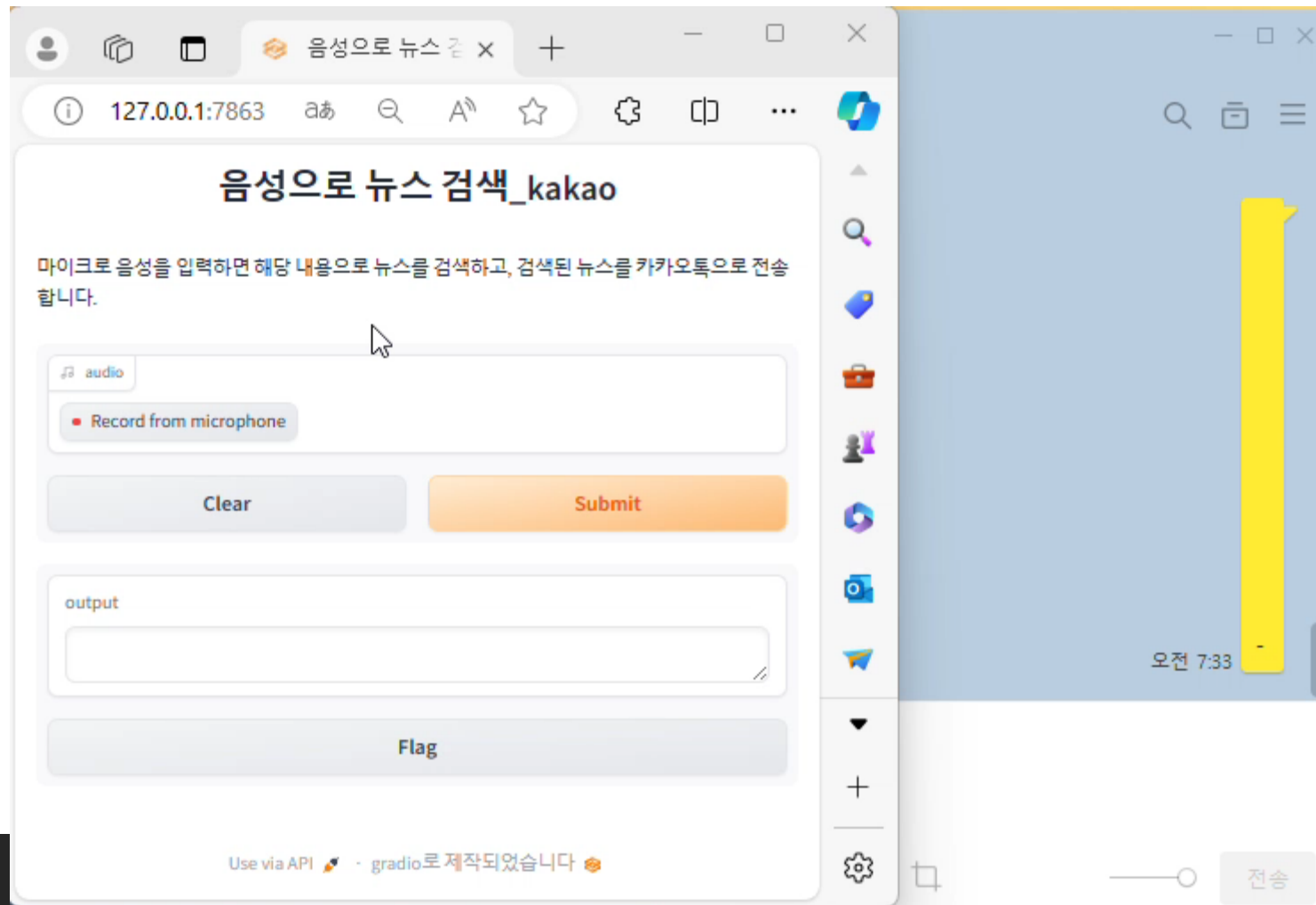
Translated Text

Flag

3. 음성 검색으로 네이버 뉴스 슬랙&카톡 메시지 전송 demo 영상 (Slack)



3. 음성 검색으로 네이버 뉴스 슬랙&카톡 메시지 전송 demo 영상 (카톡)



프로젝트 회고

프로젝트를 진행하며 느낀점

- 서경호 : STT에서 번역을 할 때 번역 Transformer 모델 부분이 더 필요한 줄 알고 있었는데 whisper 모델의 경우는 동시 번역이 가능하여 별도의 번역 모델 부분이 없어서 편리했던 것 같고 다음에는 음성 번역 뿐만 아니라 크롤링한 부분을 가지고 기능을 더 만들 수 있으면 좋을 것 같습니다.
- 기현우 : 프로젝트를 완성하는 과정에서 여러 이슈가 생겨 힘든 부분이 많았지만 처음에 기획한 주제는 전부 완성하여 뿌듯하였습니다. 추후 만약 이 프로젝트를 디벨롭 한다면 번역 모델을 따로 파인튜닝하여 STT + Translator 구조로 좀 더 성능이 좋게 만들고 싶습니다. 이번 프로젝트에서 많은 역할을 맡게 되었는데 팀 프로젝트 이다보니 확실히 팀원의 중요성을 깨닫게 되었습니다. 대학원을 다음달부터 가셔서 부트캠프를 하차하시고 팀프로젝트에 참여하지 않으셔도 되는데 시작한거 끝까지 해보시겠다고 같이 협업 해주신 경호님, 그리고 어려운 주제라 따라오시기 힘든 상황에서 본인이 할 수 있는 최선을 다해서 결과물을 만들어주신 아윤님께 감사드립니다.
- 양소현 : 상대방의 의견에 따른 근거만 듣지 말고, 상대방의 의견을 듣고 나도 그 의견의 측면에서 그 반대되는 의견에 대한 근거를 능동적으로 생각해봐야겠다. 그럼으로써 해당 의견을 더 잘 분석할 수 있을 것이다.
- 이아윤 : 수업 때 배운 내용을 적용해볼 수 있어 좋았고 카톡 채널 알림(NCP 이용)을 통해 메시지가 전송되는 기능을 추가로만 들어보고 싶습니다. 그리고 팀원분들이 프로젝트 진행하시는 모습을 보며 많이 공부해야겠다고 생각했습니다.

QnA