6주차

태그	
참석자	기흥 김Ⓟ PARKSEONGJUN 은지 하

회의 이슈

- 데이터 수집 및 처리 방안
  - 복호화 방안(mp4 → transcrition, wav)
- STT model 설계

회의 내용(문제, 현재 진행 상황, 앞으로 계획 등)

데이터 수집 및 처리 방안

- **상황:** 현재 발주사와의 계약 전
- **데이터 수집**: 도메인의 요청에 맞추어 우선 자막이 함께 있는 뉴스 영상으로부터 오디오와 전사를 추출하는 작업을 통해 dataset 구축.
- **데이터 처리**: 기존에 OSS tool을 활용해서 뉴스 영상으로부터 audio(wav or mp3)와 자막(txt or vtt)을 추출

결정 사항(내용, 진행 일정)

- **데이터 수집:** 우선 유튜브 채널 중, 미국 방송사(CNN, ABC 등)를 target으로 영상 데이터를 수집한다. 이때, 영상에는 자막이 함께 포함되어 있어야 하며, STT 엔진이 생성한 자동 생성 자막은 사용하지않는다.
- **데이터 처리**: yt-dlp라는 youtube 링크로부터 mp3, vtt 등의 파일로 영상을 복호화하는 OSS 툴을 활용하여 데이터를 복호화한다.
- **수집할 양:** 50시간 내외 분량으로 수집할 것.

특이사항

- 발주사와의 계약 전, 또는 데이터를 받기 전 까지는 팀 내에서 자체수집한 data를 활용하여 model training, test에 사용할 예정이므로 data 품질을 잘 고려하여 수집할 것