

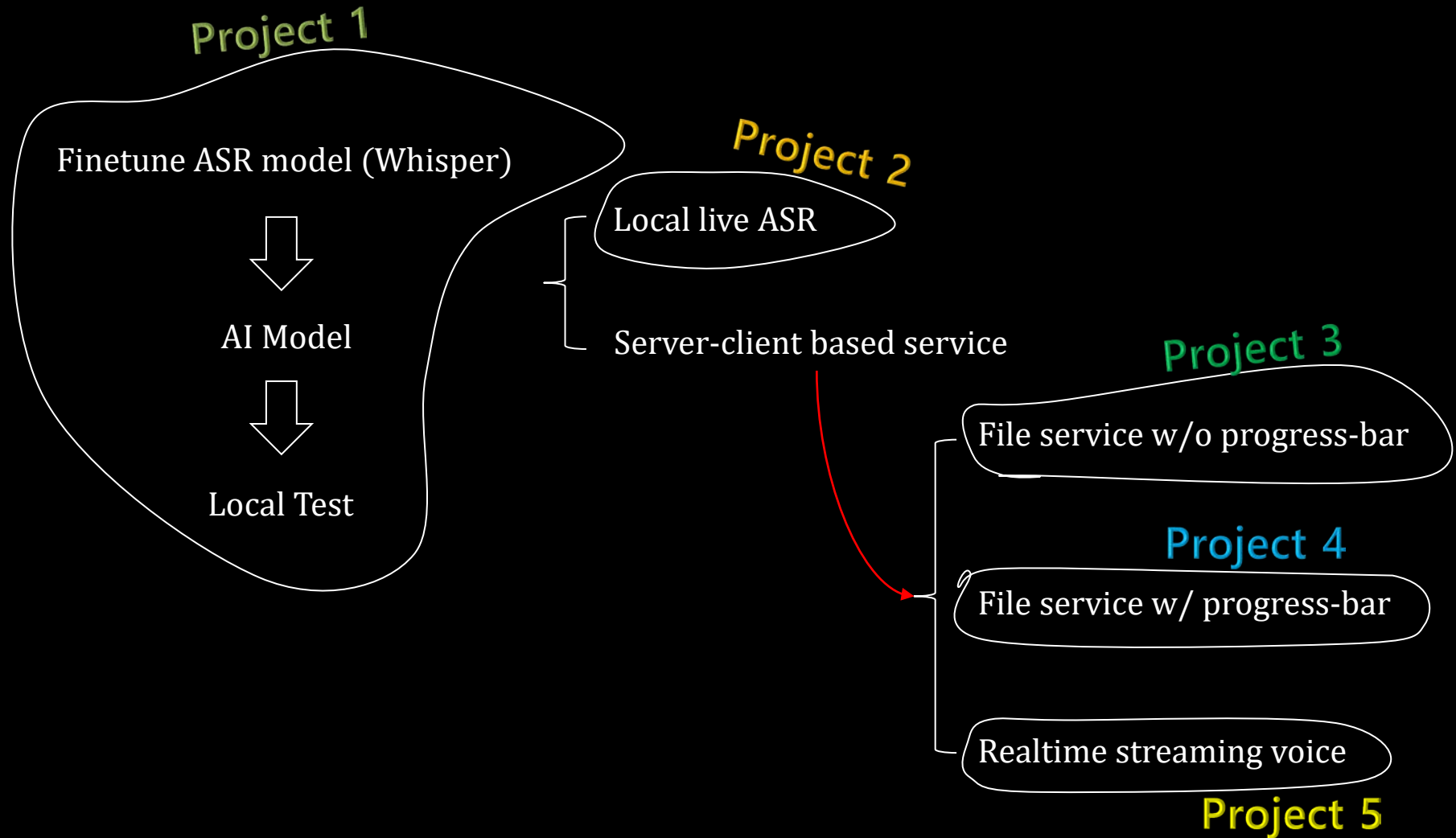
Dataset Processing

소프트웨어 끝대 강의

노기섭 교수

(kafa46@cju.ac.kr)

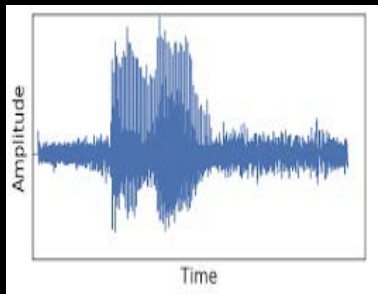
Big Picture



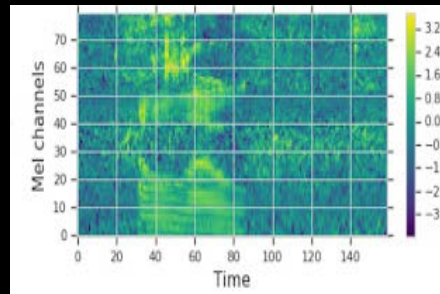
Whisper finetuning data structure

Reference: <https://huggingface.co/blog/fine-tune-whisper>

```
{'audio': {'path':  
'/home/sanchit_huggingface_co/.cache/huggingface/datasets/downloads/extracted/607848c7e74a89a3b5225c0fa5ffb947  
0e39b7f11112db614962076a847f3abf/cv-corpus-11.0-2022-09-21/hi/clips/common_voice_hi_25998259.mp3',  
'array': array([0.0000000e+00, 0.0000000e+00, 0.0000000e+00, ..., 9.6724887e-07,  
1.5334779e-06, 1.0415988e-06], dtype=float32),  
'sampling_rate': 48000},  
'sentence': 'खीर की मिठास पर गरमाई बिहार की सियासत, कुशवाहा ने दी सफाई'}
```



음성파일(mp3, wav, ...)



log-Mel spectrogram



“안녕하세요,
ACIN 아카데미 입니다.”

Label(text)

Required dataset for fine tuning

Anyway, we need tuple type data

01. (audio file, transcript)

02. (audio file, transcript)

03. (audio file, transcript)

⋮

Our target data for fine tuning

<https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=123>



The screenshot shows the AI Hub website interface. At the top, there's a navigation bar with links: AI 데이터찾기, AI 허브소개, 참여하기, 커뮤니티, AI 개발지원, 고객지원, 마이페이지, and 로그아웃. Below this, the main header says '데이터 찾기' (Find Data) with a home icon and a breadcrumb trail 'AI 데이터찾기 > 데이터 찾기'. The main content area features a large purple speaker icon with a '2' and 'L' on it. To the right of the icon, there's a title '한국어 음성' (Korean Voice) and a list of tags: #일상 대화, #쇼핑 대화, #정치 대화, #경제 대화, #취미 대화, #AI 비서, #동시통역, and #감성형 대화 음성지능 서비스. Below the title, there's a section for '분야' (Field) with '한국어' selected and '유형' (Type) with '오디오, 텍스트' selected. Further down, it shows '구축년도: 2018', '갱신년월: 2019-05', '조회수: 41,572', '다운로드: 13,800', and '용량: 72.04 GB'. At the bottom, there's a red '다운로드' (Download) button and a box with '관심데이터 취소' (Cancel Favorite Data) and a heart icon with the number '120'.

Download dataset

The image shows the AI Hub website interface for downloading the '한국어 음성' (Korean Speech) dataset. The website header includes navigation links like 'AI 데이터찾기', 'AI 허브소개', '참여하기', '커뮤니티', 'AI 개발지원', '고객지원', '마이페이지', and '로그아웃'. The main content area features a large icon of a speaker and the text '한국어 음성'. Below this, there are filters for '분야' (Korean) and '유형' (Audio, Text), and a '다운로드' (Download) button. A red arrow points from the '다운로드' button to a browser window.

The browser window, titled 'Innorix 파일 - Chrome', shows the URL: `sftp.aihub.or.kr/file/downloadList.do?dataSetSn=123&dataSe=aidata&dwldTkn=20240411...`. The page content includes the title 'AI 학습용 다운로드' and a list of instructions:

- **[+]** 를 클릭하시면 하위 폴더와 파일 목록을 확인할 수 있습니다.
- 전체 파일을 한번에 다운로드 받고자 할 경우는 [전체 다운로드] 를, 일부만 선택하여 다운로드 받고자 하실 경우는 다운로드 받을 파일을 선택하신 뒤, [선택 다운로드] 버튼을 눌러주세요.
- **주의 사항**
파일 이어받기가 안되는 경우 제어판의 INNORIX EX Agent를 삭제 후 재설치 하시길 바랍니다.
기존 설치된 Agent를 삭제하시면 install 페이지로 이동합니다.
[수동설치](#)

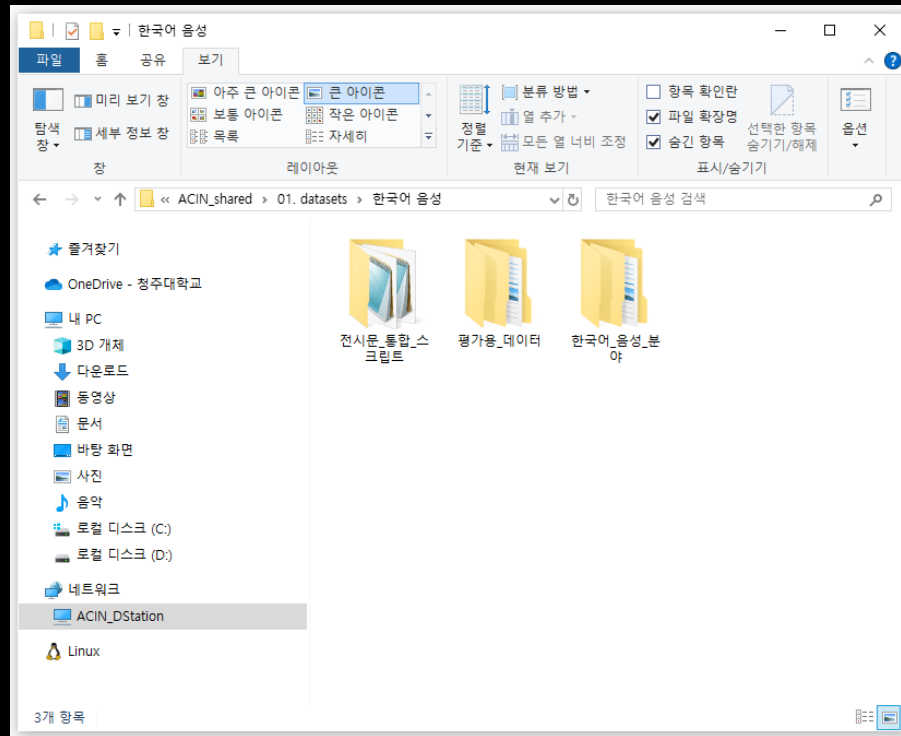
Below the instructions is a table listing the available files for download:

<input type="checkbox"/>	NAME ▲	SIZE ▲
<input type="checkbox"/>	한국어 음성	
<input type="checkbox"/>	전시문_통합_스크립트	
<input type="checkbox"/>	KsponSpeech_scripts.zip	24MB
<input type="checkbox"/>	평가용_데이터	
<input type="checkbox"/>	KsponSpeech_eval.zip	536MB
<input type="checkbox"/>	한국어_음성_분야	
<input type="checkbox"/>	KsponSpeech_03.zip	14GB
<input type="checkbox"/>	KsponSpeech_04.zip	14GB

At the bottom of the browser window, there are two buttons: '선택 다운로드' (Select Download) and '전체 다운로드' (Full Download).

Overlook dataset

Download 데이터 살펴보기



TO-DO list

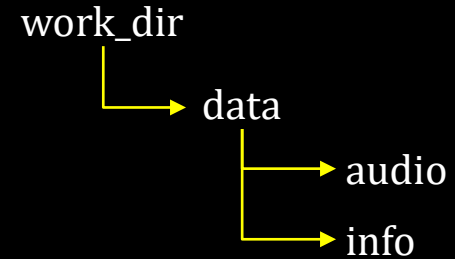
Create virtual environ & work directory

Move (or copy) files in our working directory

Convert all text files (labels or transcripts) into utf-8 encoding

Process '.pcm' files into '.wav' format

Split dataset into 'train' & 'test'





수고하셨습니다 ..^^..