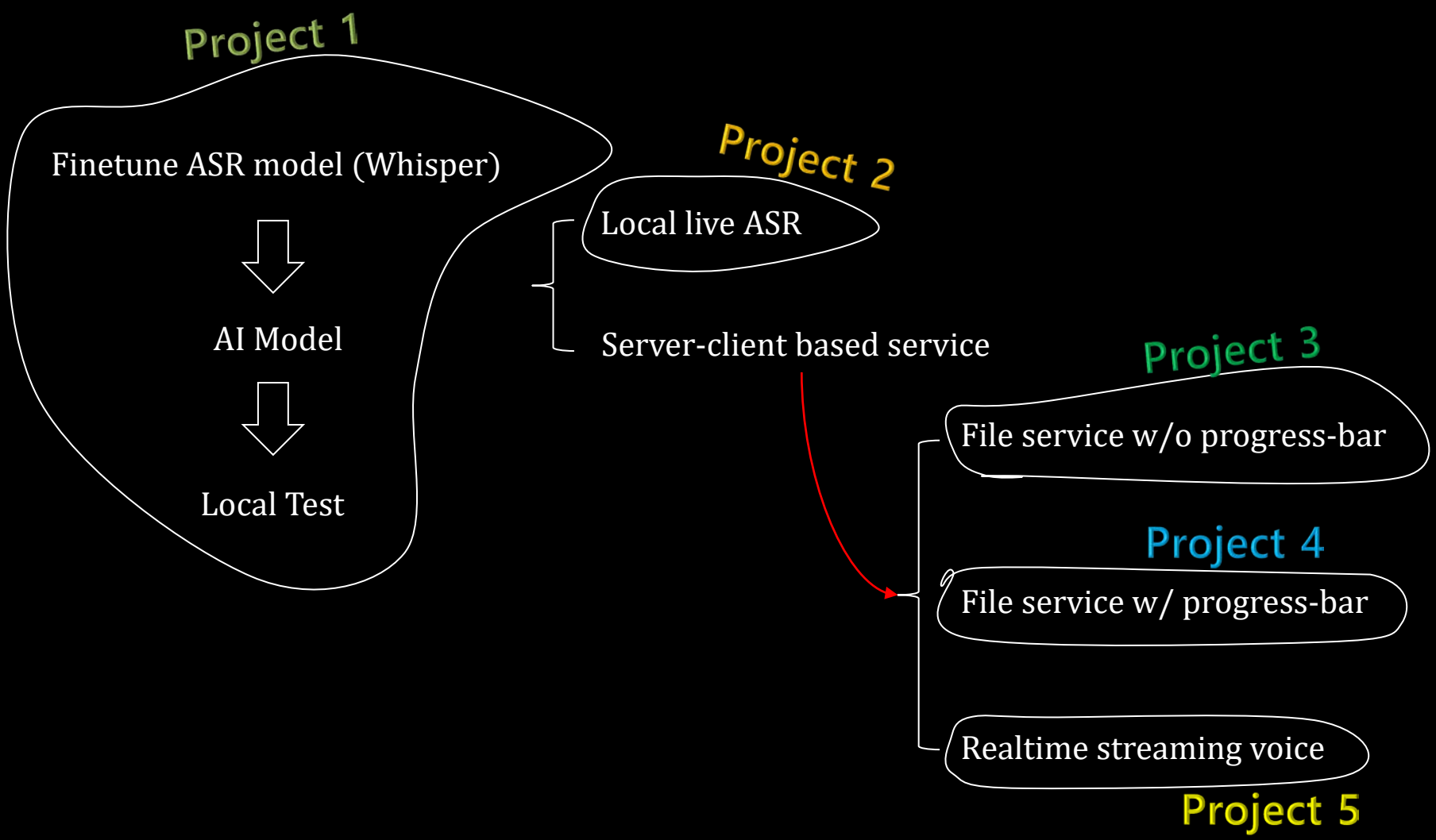
Dataset Processing

소프트웨어 꼰대 강의

노기섭 교수

(kafa46@cju.ac.kr)

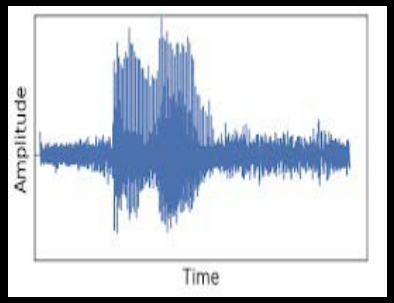
Big Picture


Project 1
Finetune ASR model (Whisper)
AI Model
Local Test
Project 2
Local live ASR
Server-client based service
Project 3
File service w/o progress-bar
Project 4
File service w/ progress-bar
Realtime streaming voice
Project 5

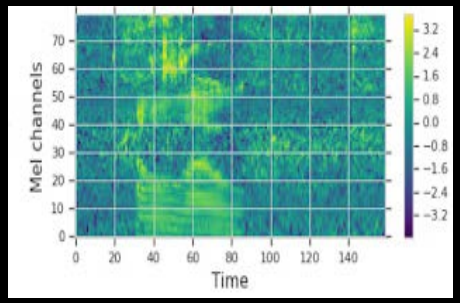
Whisper finetuning data structure

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

{'audio': {'path':
'/home/sanchit_huggingface_co/.cache/huggingface/datasets/downloads/extracted/607848c7e74a89a3b5225c0fa5ffb9470e39b7f11112db614962076a847f3abf/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)

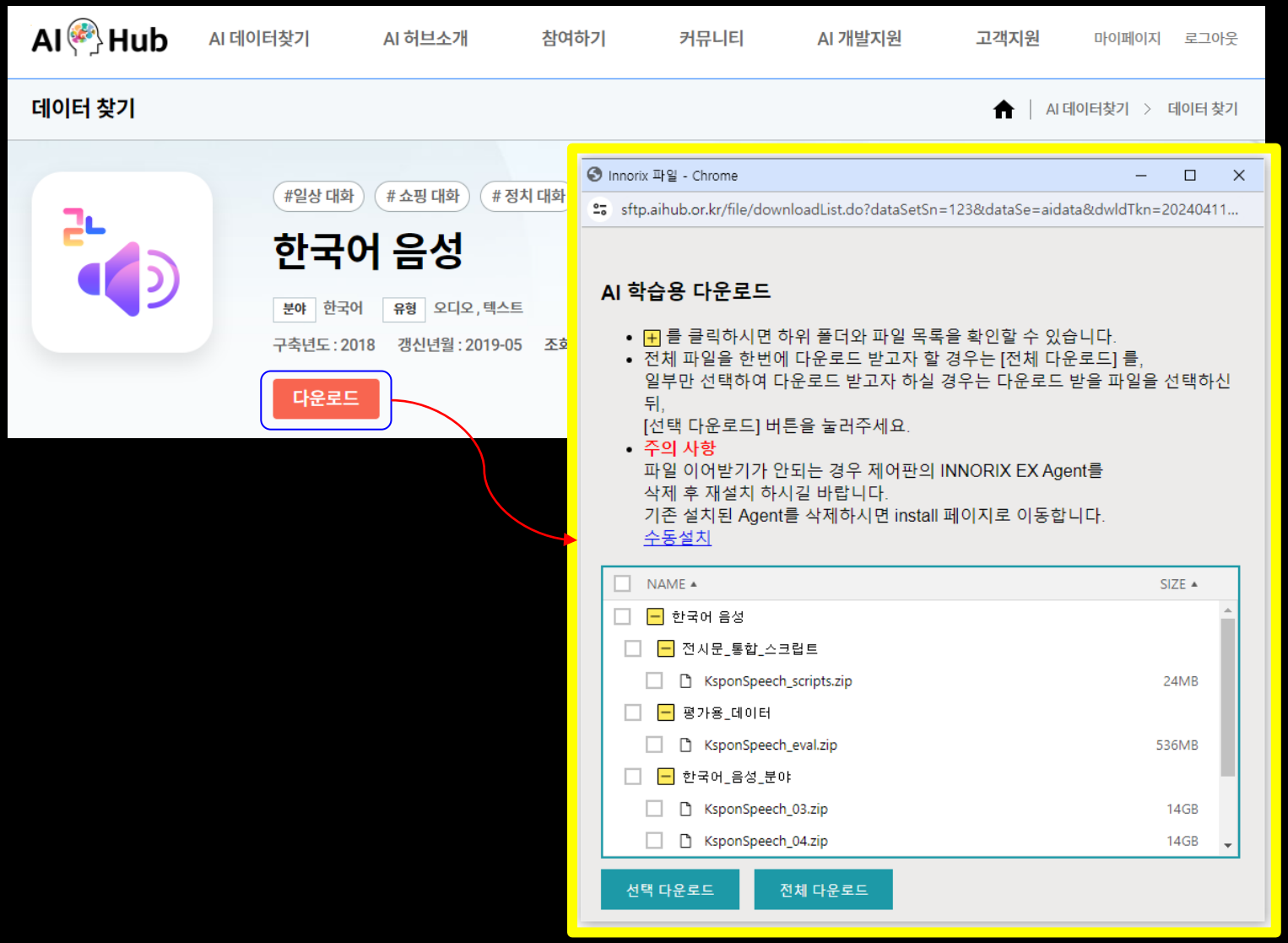
03. (audio file, transcript)

⋮

Our target data for fine tuning

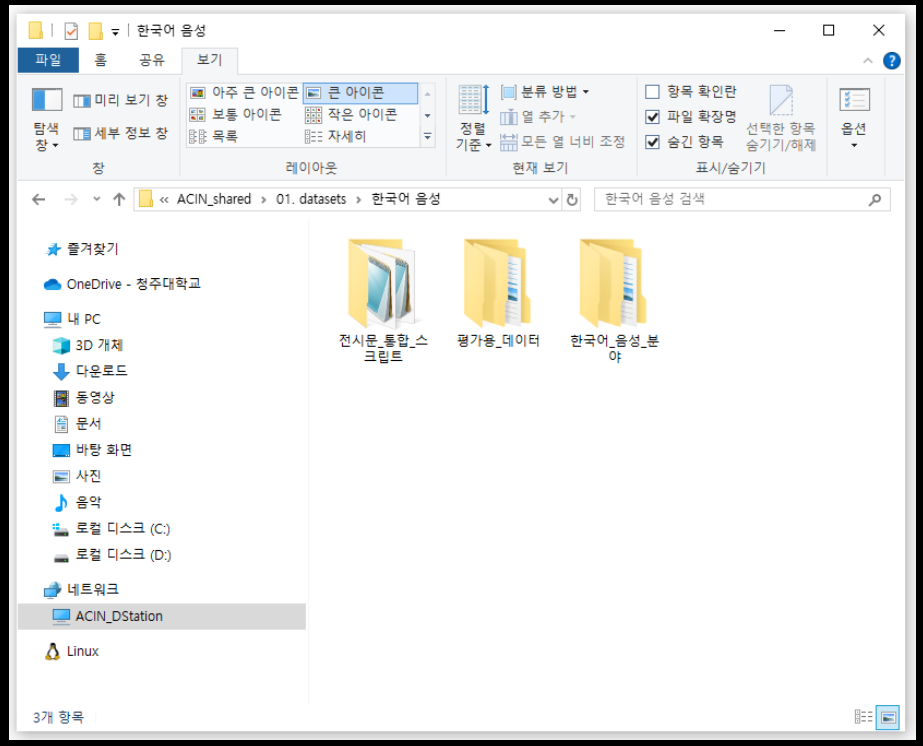
https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=123

Download dataset



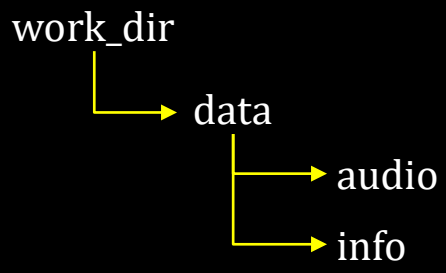
Overlook dataset

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 ..^^..