

데이터 수집

- 한류 콘텐츠에서 찾을 수 있는 일상 대화
 - 한류 연예인들이 출연하는 광고 영상
 - 한류 예능, 드라마에서의 일상 대화
 - V-Live, 인스타그램 라이브 방송 등에서의 일상 대화

+) aihub의 자유대화 음성

- 수집 방법
 - pytube를 이용한 Youtube crawling
 - request를 이용한 특정 keyword 검색 및 response json 분석

데이터 feature

발음 평가 기준은 크게 정확성 / 명확성으로 구분됨

이 프로젝트는 정확성에 중점을 두고 차후 명확성으로 확장 가능

- 정확성 : 얼마나 정확한 말로서 발음을 하는가

음성 데이터에 STT 적용하여 얻어낸 텍스트 -> 음소 sequence로 추출

- 명확성 : 발음의 길이(duration)과 억양(pitch) 등이 얼마나 정확한가

pitch contour

VAD(Speech detection)을 이용해 얻어내는 duration