

- 자유대화 과제

## 자유대화(일반남여)



### ●○ 개요: 자연어 자유대화(일반남여) 학습데이터셋이란?

자연어 자유대화(일반남여) 학습데이터셋은 일반남여 사용자들의 자유대화 음성데이터 및 음성데이터 전사결과, 녹음 대상자의 정보, 녹음환경 등의 정보를 포함한 메타데이터로 구성된다.

본 데이터를 통해 음성인식, 음성언어처리, 자연어처리, 한국어 음성언어연구, 신호처리 등의 연구 분야와 온/오프라인 기반의 음성인식, AI비서, Voice Bot, Voice Command & Control, AI 로봇, 음성인식 기반 키오스크 등의 산업분야에 활용할 수 있다.



그림1 | 자연어 자유대화 학습데이터셋 구축

### ●○ 데이터셋의 구성

자연어 자유대화(일반남여) 학습데이터셋은 아래와 같이 구성된다.

|                | 데이터 요약데이터 구분  | 제공 방식  |
|----------------|---|--|
| 자유대화<br>(일반남여) | - 10대에서 50대 사이의 일반인 남녀의 발화 데이터<br>- 녹음 인원 2,000명 이상, 4,000시간 음성 데이터 | - 원천데이터: PCM(WAV) 파일 포맷<br>- 메타데이터: JSON 파일 포맷 |

### ●○ 데이터셋의 설계 기준과 분포

본 데이터셋은 성별, 연령, 수집방법, 구축량에 따라 각각의 설계기준을 제시하며, 설계 기준에 따라 성별, 연령, 장소, 지역 환경에 따라 구축 데이터가 분포된다.

## ●○ 데이터 설계 기준

| 과제명                    | 주요 내용                               | 데이터 수집 방법               | 데이터 구축량                | 데이터 형식                                |
|------------------------|-------------------------------------|-------------------------|------------------------|---------------------------------------|
| 데이터1<br>자유대화<br>(일반남여) | 10~50대의 화자의<br>음성 데이터<br>(남녀비율 1:1) | 오프라인(스튜디오)<br>온라인(음성채팅) | 2000명 이상의 화자<br>4000시간 | 음성데이터/<br>텍스트데이터<br>(음성과 매칭)/<br>관련정보 |

## ●○ 데이터 분포

- 유형별 객체 클래스 분포
  - 일반남여: 2,000명 화자 및 4,000시간 분량의 음성 데이터
- 성별, 연령별 분포
  - 일반남여 남녀 비율은 1:1 (비율의 차이는 10%미만)
  - 일반남여는 10~50대를 포함
- 성별 분포(남여 비율 1:1 (비율의 차이는 10%미만))

|                   |   |               |   |               |
|-------------------|---|---------------|---|---------------|
| 일반남여 합계<br>2,000명 | = | 남자<br>약1,000명 | + | 여자<br>약1,000명 |
| 100%              |   | 약 50%         |   | 약 50%         |

- 장소, 지역 환경 분포
  - 지역에 따른 억양 사투리 및 단어 사투리를 반영해야 하므로 전국 지역의 분포 확인
  - 수집방법에 따라 시챗봇, 음성수집도구, 스튜디오 분포 확인

## ●○ 데이터 구조

본 데이터는 원천 및 메타 데이터로 구분되며, 원천 데이터는 대상자 및 대화 시나리오 정보를 포함한 음성파일을 의미하고 메타데이터는 대상자의 정보, 음성데이터의 전사 결과(음성인식 결과), 녹음환경 및 대화 주제 정보를 포함한 데이터를 말한다.

### ● 데이터 수집/가공 형태

| 수집 대상 | 형태  |
|-------|---|
| 원천데이터 | <ul style="list-style-type: none"> <li>PCM(WAV) 음성 파일</li> <li>대상자 및 대화 시나리오 정보를 포함한 음성파일</li> </ul>  |
| 메타데이터 | <ul style="list-style-type: none"> <li>Json 형태</li> <li>대상자 정보 (성별/연령/지역)</li> <li>녹음환경 정보 (실내/실외: 대중교통, 거리 등)</li> <li>대화 주제 및 음성데이터 전사결과</li> </ul> |

※ 원천데이터(음성파일)와 메타데이터(Json)로 구분

※ 원천데이터(음성파일)은 각각의 파일명으로 구분 (Ex. sample1.wav)

### ● 메타데이터 항목

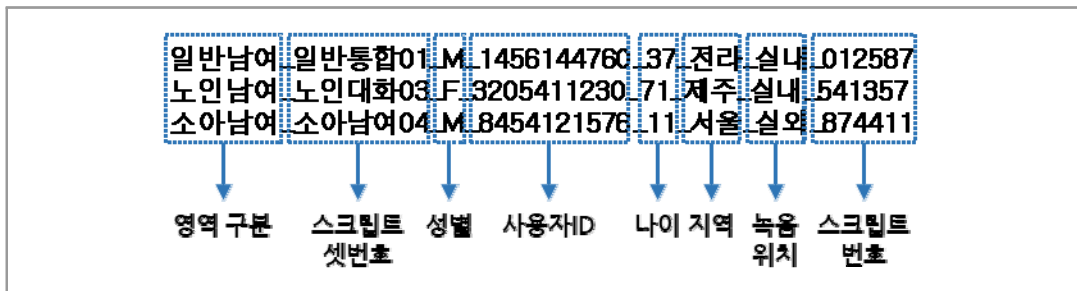
| 대분류   | 속성표기          | 의미       | 타입     | 필수여부 |
|-------|---------------|----------|--------|------|
| 대화정보  | colctUnitCode | 수집방법     | String | Y    |
|       | convrsThema   | 대화주제     | String | Y    |
|       | cityCode      | 지역       | String | Y    |
|       | recrdEnvrn    | 녹음환경     | String | Y    |
|       | recrdUnit     | 녹음도구     | String | Y    |
| 녹음자정보 | recorder      | 녹음자      | String | Y    |
|       | recorderId    | 녹음자ID    | String | .    |
|       | gender        | 성별       | String | Y    |
|       | age           | 나이       | String | Y    |
| 발화정보  | recrdDt       | 녹음일시     | String | Y    |
|       | recrdTime     | 녹음시간     | String | Y    |
|       | stt           | 음성인식결과   | String | Y    |
|       | fileNm        | 파일명      | String | Y    |
|       | recrdQuality  | 녹음품질     | String | Y    |
|       | scriptSetNo   | 스크립트셋 번호 | String | .    |
|       | scriptId      | 스크립트ID   | String | .    |

• 메타데이터 항목별 예시와 규격

- 파일 명명 규칙에 따르며 파일 명명 규칙은 '영역구분\_스크립트셋번호\_성별\_사용자ID\_나이\_지역\_녹음위치\_스크립트번호'에 따름.
  - Ex) 일반남여\_일반통합01\_M\_1456144760\_37\_전라\_실내\_012587
- 스크립트번호 : 대화스크립트번호(연번 내 번호만)
- 영역구분 : 일반남여 / 노인남여 / 소아남여 / 외래어
- 스크립트셋번호 : 일반통합00 / 노인대화00 / 소아남여00 / 외래어00
- 성별 : M(남성) / F(여성)
- 사용자ID : 각 녹음 사용자 식별 값
- 나이 : 숫자(정수)
- 지역 : 수도권 / 충청 / 강원 / 전라 / 제주 /기타
- 녹음위치 : 실내 / 실외 / 녹음실
- 스크립트번호 : 숫자(정수)

| 영역 구분 | 스크립트 셋번호 | 성별    | 사용자ID  | 나이     | 지역  | 녹음위치 | 스크립트 번호 |
|-------|----------|-------|--------|--------|-----|------|---------|
| 일반남여  | 일반통합00   | M(남성) | 사용자식별값 | 숫자(정수) | 수도권 | 실내   | 숫자(정수)  |
| 노인남여  | 노인대화00   | F(여성) |        |        | 충청  | 실외   |         |
| 소아남여  | 소아남여00   |       |        |        | 강원  | 녹음실  |         |
| 외래어   | 외래어00    |       |        |        | 전라  |      |         |
|       |          |       |        |        | 제주  |      |         |
|       |          |       |        |        | 기타  |      |         |

- 저장 시 아래 사례와 같은 파일 이름으로 저장이 가능하며, 파일명으로 데이터 구조 파악 가능



## ●○ 데이터 예시

메타 데이터는 Json형태로 각 속성과 값으로 쌍을 이루는 구조를 가진다.

- 메타데이터 형태(Json)

```
{
  "대화정보":{
    "colctUnitCode":"음성수집 도구"
    "convrsThema":"국내 여행"
    "cityCode":"수도권"
    "recrdEnvrn":"집안"
    "recrdUnit":"ANDROID"
  },
  "녹음자정보":{
    "recorder":"홍길동"
    "recorderId":"HONGKD"
    "gender":"남"
    "age":"20"
  },
  "발화정보":{
    "recrdDt":"2020-08-28 12:31:00"
    "recrdTime":"5.13"
    "stt":"한산도 해안도로는 섬들을 보며 바닷가를 달리는 해안드라이브 코스입니다."
    "fileNm":"sample-1.wav"
    "recrdQuality":"16K"
    "scriptSetNo":"일반남녀01"
    "scriptId":"일반남녀-00001"
  }
}
```

## ●○ 데이터 구축 과정

데이터 구축 과정은 수집, 가공, 검수, 학습, 응용 단계로 수행되며, 각 단계별 수행내역은 다음과 같다.

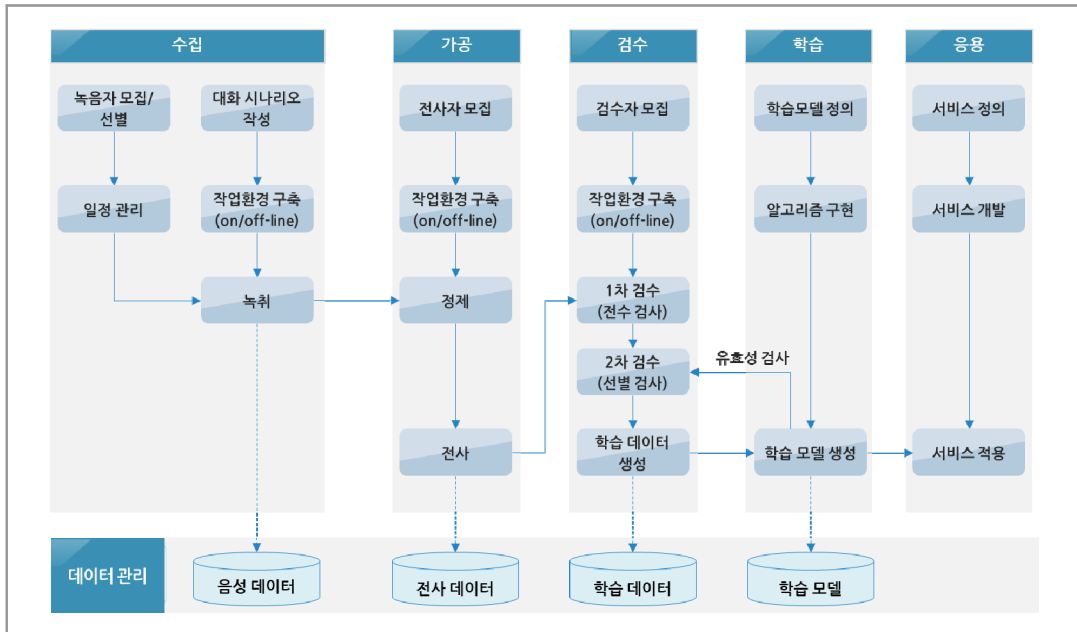


그림2 | 데이터 구축 개요

- 수집
  - 음성 데이터를 녹음할 녹음자를 모집 및 선별하고 작업 일정을 관리
  - 녹음자가 녹음 할 대화 시나리오(스크립트)를 작성
  - 온라인 / 오프라인 작업장을 구축하고 녹음자와 일정을 협의하여 녹취 진행하고 음성데이터 생성
  - 입력되는 음성 데이터 샘플링 주파수를 16kHz로 통일하여 설정하고, 44kHz 샘플링 rate로 녹음을 진행하는 경우 16kHz로 다운 샘플링 처리
- 가공
  - 데이터 가공을 위한 전사자 인력 모집 및 작업 환경 확보
  - 수집 과정에서 녹취한 음성데이터를 전사자가 정제 및 전사 수행하고 최종 전사 데이터 생성
- 검수
  - 데이터 검수를 위한 검수자를 인력 확보와 검수 규칙의 정비
  - 가공 과정에서 전사한 전사데이터를 검수자가 검수(1/2차에 걸쳐서 수행)를 하고 최종 학습 데이터를 생성

- 학습
  - 학습을 위한 학습모델을 정의하고 해당 알고리즘을 구현
  - 검수 과정에서 생성한 학습 데이터와 구현한 알고리즘을 바탕으로 학습 모델을 생성
  - 수집된 DB를 통해 BASE 엔진 기반으로 각 분야별, 음향, 언어모델 적용 학습을 수행하고, BASE 엔진 대비 인식 성능향상 여부의 유효성 검증을 진행
  - 각 분야별 응용 서비스에도 인식엔진을 제공하므로, 수요 업체의 서비스 시나리오의 유즈 케이스 문장을 통해 언어모델 적용 학습을 하여 언어모델 적용 학습 진행
- 응용 (응용 서비스 구성 시)
  - 학습데이터를 기반으로 한 응용 서비스를 정의
  - 정의된 서비스를 바탕으로 서비스를 개발하고 학습 모델이 생성되면 해당 서비스를 적용하여 응용서비스 개발

●○ 검수와 품질 확보

검수는 수집, 정제/가공, 검수, 학습/검증 단계로 진행된다. 원천데이터와 어노테이션 작업 결과를 수집한 뒤에 음성 데이터 가공, 전사화 전수 검수, 어노테이션 가공, 데이터셋 검수를 수행하고 1차 전수 검사와 유효성 검사 및 2차 선별 검사 결과를 음성인식 모델에 학습하여 검증하는 단계로 진행된다.

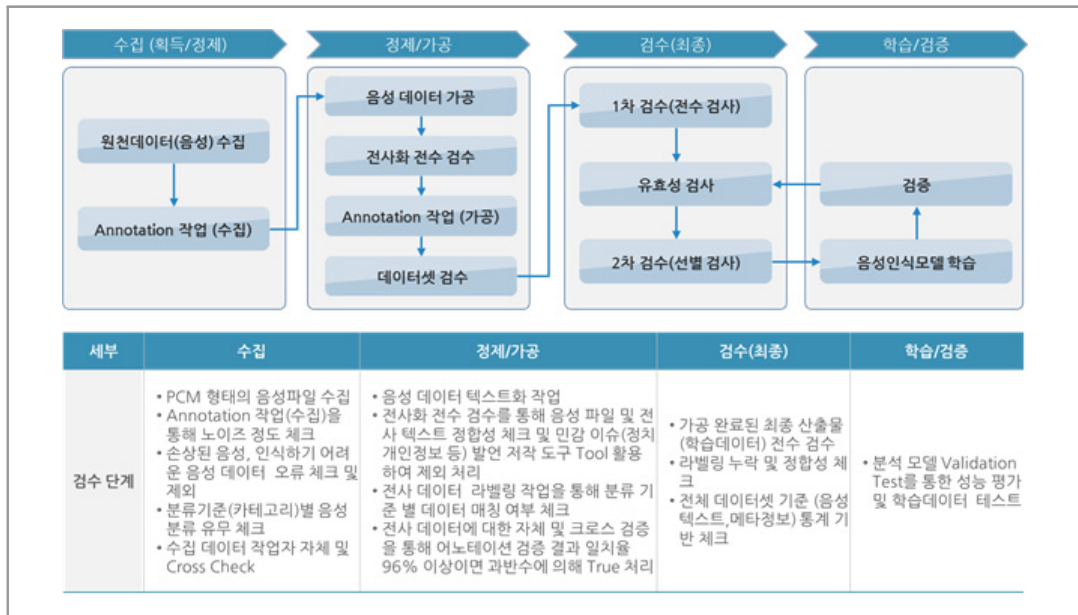


그림3 | 품질관리 프로세스

검수는 작업자, 관리자 별 검수와 최종 검수의 3단계 과정을 통해 이루어지며, 검수단계에서는 크로스 체크를 통해 검수 일관성을 유지하게 된다.

| 검수 단계             | 세부 내용  |
|-------------------|--|
| 작업자 검수 (수집)       | <ul style="list-style-type: none"> <li>작업자 스스로 리뷰를 진행한다.</li> </ul>  |
| 관리자 검수 (수집/정제/가공) | <ul style="list-style-type: none"> <li>작업자 - 관리자간의 리뷰를 진행한다.</li> <li>품질 기준에 위배되는 데이터는 재작업을 진행한다.</li> <li>범위 작업 : 데이터 수집(환경정보 수집데이터 체크)<br/>데이터 라벨링 작업<br/>데이터 전사화 작업<br/>데이터 민감정보 삭제 작업</li> </ul> |
| 최종 검수             | <ul style="list-style-type: none"> <li>데이터 라벨링 확인</li> <li>데이터 어노테이션 누락여부 확인</li> <li>학습데이터 정확도 확인</li> <li>민감정보 포함 유무 확인</li> </ul>   |

### ●○ 데이터 구축 담당자

수행기관(참여기관) : NHN다이퀘스트

(전화: 02-3470-4307/4374, 070-4658-4425,

이메일: help2.workpedia@diquest.com