●○ 명령어 과제

명령어 데이터(노인남여/정형·비정형 포함)

●○ 개요: 명령어 AI 데이터셋이란?

본 프로젝트는 명령어 AI 데이터를 수집하는 것이 주목적이며, 궁극적으로는 딥러닝 기반 STT(Speech To Text) 엔진의 성능을 개선하는 것이다. 따라서 본 프로젝트 구축 데이터는 녹취 음성 파일 형식으로 구축된다.

본 구축 데이터를 통해 한국인의 다양한 환경에서의 음성인식 결과를 활용한 서비스 및 제품을 개발하려는 기업, 연구기관, 연구자 등에게 기술 개발의 리소스를 제공하며, 나아가 AI 스피커, AI 로봇, 키오스크, 스마트카 등 음성인식과 관련된 미래 사업에 AI 기술 활용도를 높이는데 기여할 수 있다.

본 데이터 구축은 시나리오 설계, 데이터 수집, 데이터 태깅, 데이터 검수의 과정으로 진행되며, 중간 결과물을 통한 STT 엔진 모델링을 수행하여 모델 성능을 확인한다. 이어 대규모 데이터 구축을 수행한 뒤 공인 인증 기관의 품질 검증 과정을 거쳐 구축 데이터 및 STT API를 공개한다.

인공지능 학습을 위한 명령어 AI 데이터셋은 STT(Speech to Text) 기술을 활용하여 AI 스피커, AI 로봇, AI 키오스크, 차량 AI 기술 등 서비스 기능을 호출하거나 검색하는 데 사용되는 명령을 잘 인식할 수 있게 하는 원천 음성 및 전사 데이터를 말한다.

명령어 AI 데이터셋을 활용한 인공지능 모델을 사용함으로써, 사람이 말하는 음성을 텍스트로 표현하여 높은 수준의 의미 분석 처리 및 AI 음성 챗봇 등의 대화 처리가 가능하도록 한다.

명령어 AI 기술이 주로 사용되는 분야를 노년층에 해당하는 명령어 음성을 수집하고, 관련된 메타데이터를 기록한다.

명령어 AI 데이터셋의 사례는 아래와 같다.

항목	시간규모	목적
명령어 AI 데이터 (노인)	3,000	1. AI 비서
		2. AI 로봇
		3. 음성인식 키오스크

그림1 | 명령어 AI 데이터 구분

●○ 데이터셋의 구성

본 데이터셋은 명령어 AI 데이터(노년층)에 대해 대본을 생성하고, 해당 대본을 음성으로 읽어서 녹음한 음성 파일 3,000시간 분량으로 구성되어 있다. 이에 따른 전사 텍스트 데이터는 JSON 형식으로 음성과 쌍을 이루는데, 이 전사 텍스트 안에 각종 메타 데이터를 기록한다. 명령어 데이터셋 3,000시간은 연구자가 일반적으로 연구를 진행하기에 충분한 양이며, 상용화 수준에서는 강력한 STT용 음향 모델을 만들 수 있는 양이다.

한국어 음성인식 기술은 최근 다양한 서비스에서 매우 활발하게 사용되고 있다. 주요 서비스 대상은 AI 스피커, AI 로봇, 키오스크, 차량 비서 등이며, 음성인식 엔진(STT)의 활용이 기본적으로 이루어지고 있다.

노년층을 대상으로 구축한 데이터는 AI 음성비서, AI 로봇, 음성인식 키오스크 등에 활용될 수 있다.

음성인식에서 데이터의 중요성은 널리 알려진 바이다. 음성인식 엔진의 성능을 개선하기 위한 음향 모델(Acoustic Model)의 제작에서 음성 파일은 필수적으로 확보되어야 한다.

특히 노인의 음성을 취득하기는 더욱 어려운 일이다. 본 프로젝트에서는 노인의 음성을 대규모로 수집하여 STT 엔진 성능 개선에 활용할 수 있는 명령어 관련 AI 음성 데이터를 구축하고자 한다.

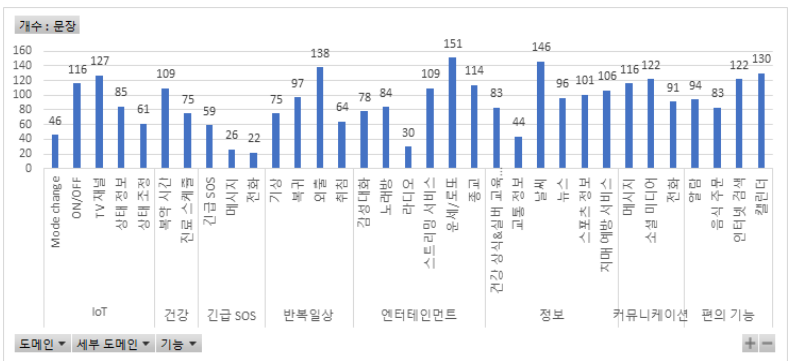
데이터 종류	포함 내용	제공 방식
음성데이터	노년층의 AI 데이터	WAV포맷 파일
전사데이터	음성 파일에 대한 전사 텍스트 및 메타 데이터	JSON 포맷 파일

●○ 데이터셋의 설계 기준과 분포

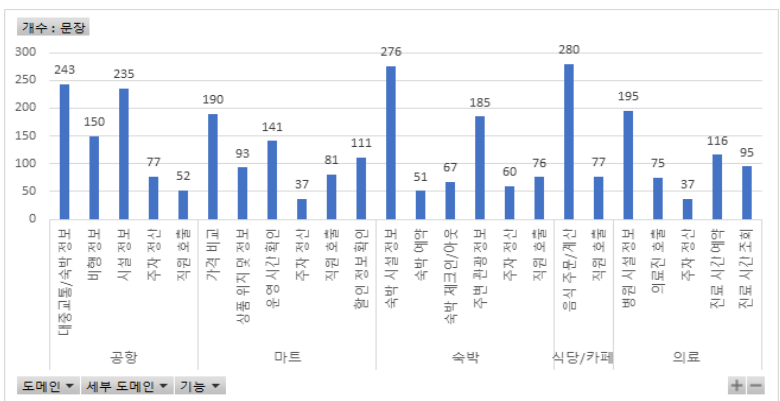
데이터셋을 설계할 때 가장 중요하게 고려했던 점은 데이터 밸런스이다. 본문과 질문, 정답을 구성할 때 적절한 분류기준을 만들었고, 해당 분류기준에 따라 골고루 데이터가 분포되도록 설계하여 학습 시 예상할 수 있는 데이터 편향성을 최소화하도록 했다.

명령어 AI 데이터는 아래와 같은 분포로 구성된다.

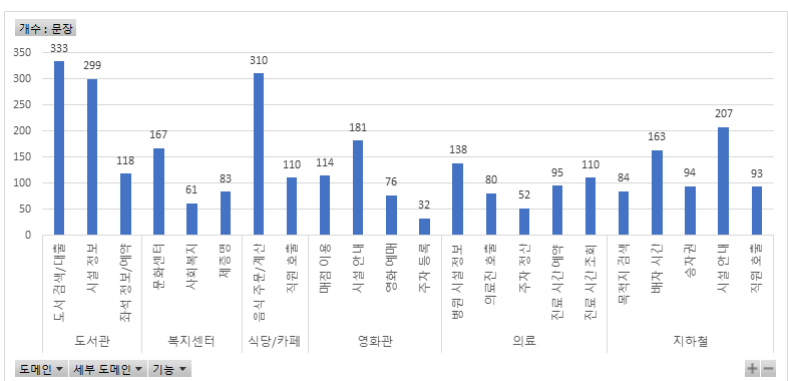
(1) 노년층 AI 비서 분포



(2) 노년층 AI 로봇 분포

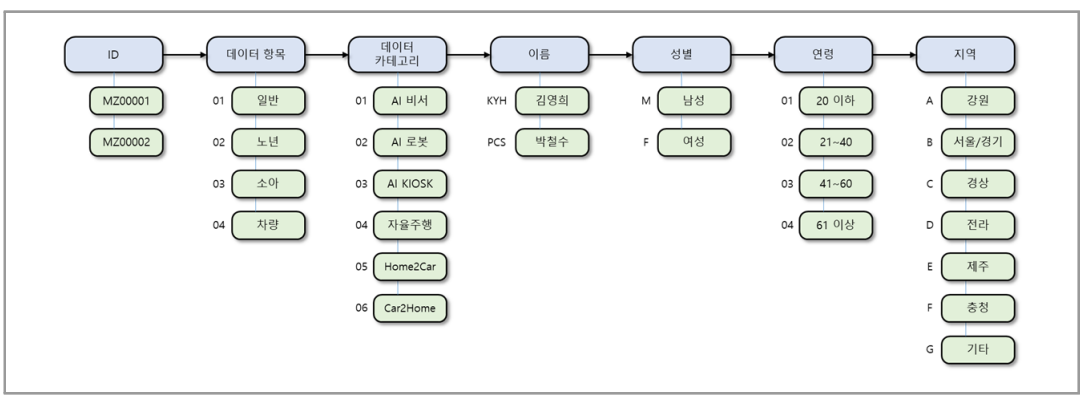


(3) 노년층 AI 키오스크 분포



●○ 데이터 구조

- 명령어 AI 데이터 파일은 아래의 구조를 따른다.
  - Wave 파일 어노테이션에 필요한 정보를 데이터 수집 단계에서 입수한 뒤, 각 항목 앞에 정의된 숫자 또는 영문 표기를 파일 명으로 저장한다.
  - 추후 해당 메타데이터는 모두 어노테이션 정보를 담고 있는 JSON 형식에 기재되어 저장된다.
  - 이때 화자 이름과 같은 개인 식별 정보는 이니셜 등으로 변경 저장한다.
  - 서버 저장 시 아래 사례와 같이 간소한 형식의 파일 이름 저장이 가능하며, 추후 메타 데이터에 일괄 정보 기입이 가능하다.



●○ 데이터 예시

음성데이터는 아래의 구조를 가진다.

(1) Wave(PCM) 데이터 포맷 : 48kHz, 16bit, mono

- 추후 16kHz 다운 샘플링이 가능한 형식으로 녹음 취득.

(2) json 데이터 포맷 : 일반 json 형식

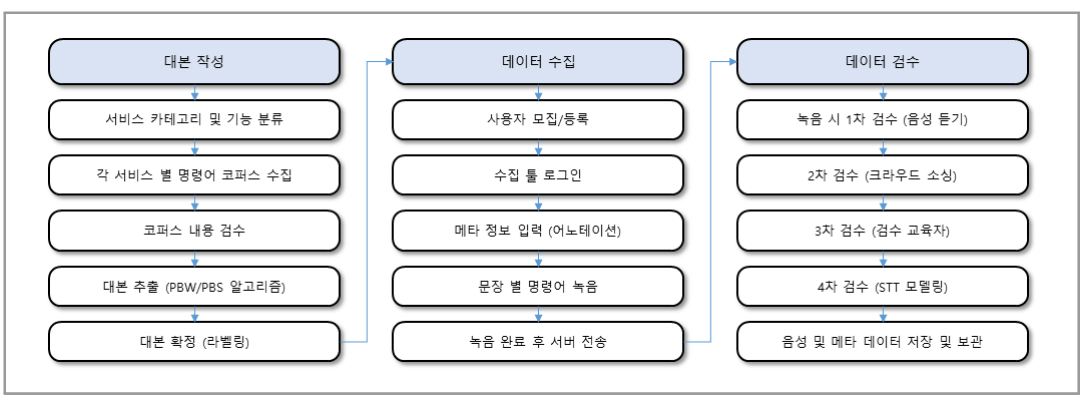
| json 형식 샘플 |


```
{
"기본정보":{
        "Language" : "ENG",
        "Version" : "N/A",
        "ApplicationCategory" : "N/A",
        "NumberOfSpeaker" : "2484",
        "NumberOfUtterance" : "N/A",
        "DataCategory" : "readSpeech",
        "RecordingDate" : "N/A",
        "FillingDate" : "N/A",
        "RevisionHistory" : "N/A",
        "Distributor" : "MediaZen"
        },
"음성정보":{
        "SamplingRate" : "48000",
        "NumberOfByte" : "16",
        "ByteOrder" : "N/A",
        "EncodingLaw" : "SignedIntegerPCM",
        "NumberOfBit" : "N/A",
        "NumberOfChannel" : "1",
        "SignalToNoiseRatio" : "N/A"
        },
"전사정보":{
        "LabelText" : "예 그렇습니다",
        },
"화자정보":{
        "SpeakerName" : "PCS",
        "Gender" : "Male",
        "Age" : "21~40",
        "Region" : "01",
        "Dialect" : "NotProvdied"
        },
"환경정보":{
        "RecordingEnviron" : "Mart",
        "NoiseEnviron" : "Mart",
        "RecordingDevice" : "SmartPhone",
        },
"파일정보":{
        "FileCategory" : "Audio",
        "FileName" : "test1_1_01_01_PCS_M_02_B.wav",
        "DirectoryPath" : "/path/to/the/folder",
        "HeaderSize" : "N/A",
        "FileLength" : "N/A",
        "FileFormat" : "PCM",
        "NumberOfRepeat" : "1",
```

```
            "TimeInterval" : "N/A",
            "Distance" : "50"
            },
"기타정보":{
            "QualityStatus" : "Good",
            }
}
```


●○ 데이터 구축 과정

데이터 구축은 다음의 과정을 거쳐서 수행된다.



데이터 구축은 AI 비서, AI 로봇, AI 키오스크 등의 도메인에 대해 각 기능 별 명령어 문장을 수집하고, 노년층으로부터 수집된 코퍼스에서 PBW(Phonetically Balanced Word) 방식으로 음소 연쇄 균형 알고리즘으로 대본을 추출한다.

아래 표는 항목 별 대본 추출의 모집단과 추출 대본의 수이다.

구분	항목	모집단 (문장)	추출 대본 수 (문장)
노년층 명령어 AI 데이터	AI 비서	32,568	3,000
	AI 로봇	24,486	3,000
	AI 키오스크	25,534	3,000
총 계		82,588	9,000

수집된 데이터는 정제/가공을 거쳐 검수를 진행한다.

획득 및 정제의 기준은 다음과 같다.

(1) 음성 품질 획득 기준
- 소음이 거의 없는 음성을 중심으로 수집
- 낮은 수준의 에코, 작은 숨소리는 허용
- 음성 앞뒤의 묵음 구간이 100ms보다 길어야 함
- 발음이 올바르고 이해할 수 있을 정도여야 함
- 녹음 상태가 양호해야 하며 전기적 노이즈나 원인 불상의 소음이 섞이지 않아야 함
- 단어와 단어 사이의 묵음이 2초를 넘지 않아야 함

(2) 음성 정제/삭제 대상
- 물체가 부딪히는 큰 소리 (책상, 의자 등)
- 사이렌이나 경적 소리
- 발자국 소리, 노크 소리
- 화자 이외의 다른 사람의 목소리
- 기침 소리 또는 큰 숨소리
- 과도한 분절이 있을 때 (너무 끊어서 발음할 경우)
- 목소리에 과도한 떨림이 있는 경우
- 침을 크게 삼키는 소리가 포함된 경우
- 음성을 우물거리며 발음할 때
- 말을 더듬으며 발음할 때
- 발음에 이상이 있을 때
- 녹음 환경 대상과 무관한 배경 소음이 지속적으로 발생할 때

(3) 음성 데이터 정제/검수 주의 대상
- 의미가 달라질 수 있는 발음 (교통 -> 고통)
- 의미는 동일하지만 받침/조사 등에 오류가 있는 경우 (끄라니까 -> 끄라니깐)
- 의미는 동일하지만 받침/조사 등의 누락 (눈이 감겨 -> 눈 감겨)
- 발음 부정확 (어눌함, 부자연스러움, 버벅거림)
- 발음 불명확 (주 -> 주부)
- 녹음 음량이 너무 크거나 작은 경우
- 녹음 속도가 너무 빠르거나 느린 경우

데이터 수집은 다양한 권역별, 연령별 피험자를 모집하여 수행한다.

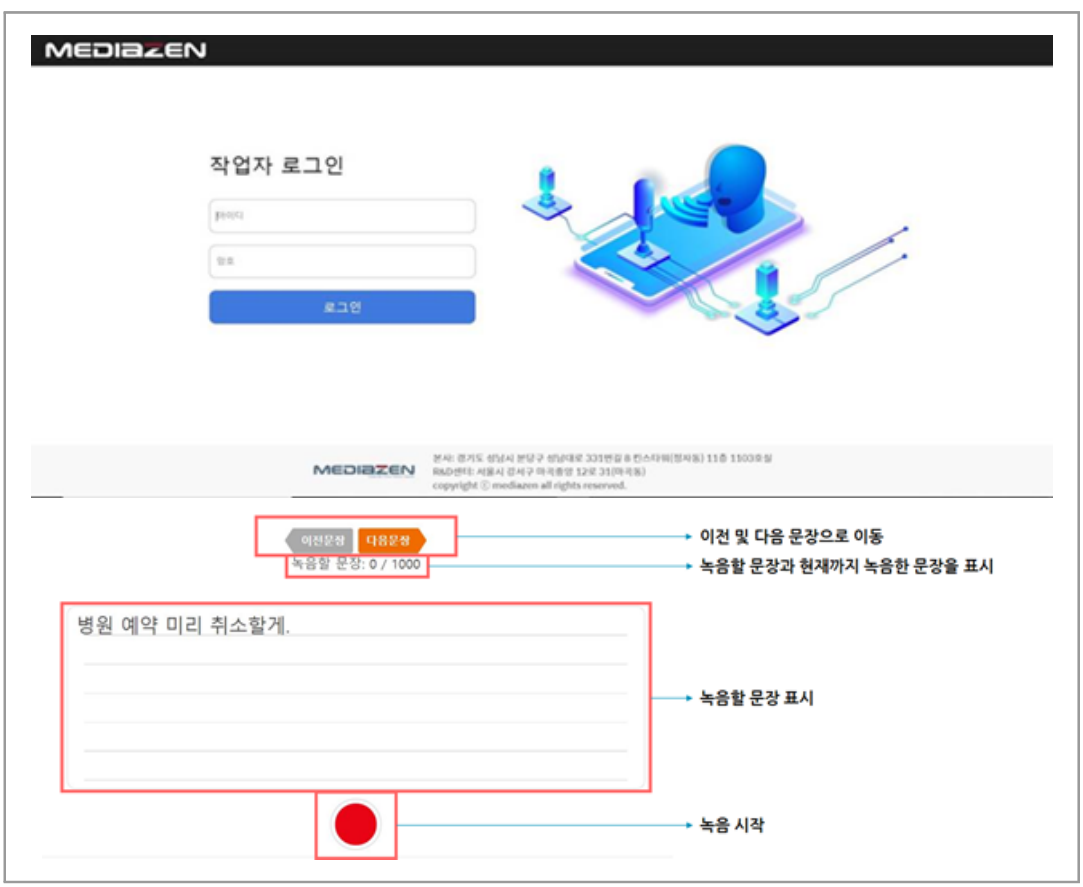
표 | 정형 권역, 연령, 성별 피험자 정보

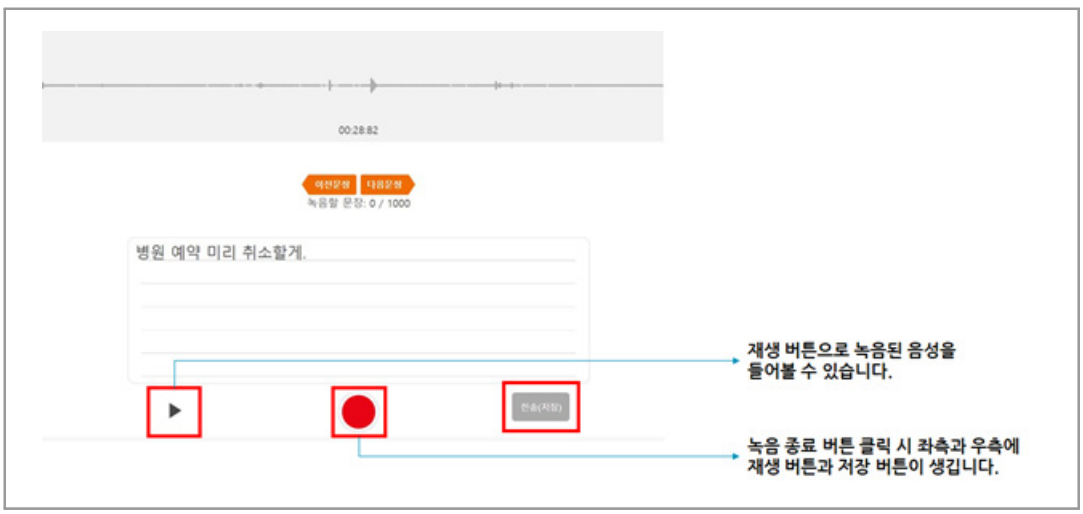
권역	노인				합계	
	60~69세		70~79세			
	남	여	남	여	남	여
서울/인천/경기	359	358	191	191	550	549
대전/세종/충청/강원	109	109	62	62	171	171
광주/전라/제주	88	89	59	58	147	147
부산/대구/울산/경상	215	215	117	118	332	333
전국	771	771	429	429	1200	1200

표 | 비정형 권역, 연령, 성별 피험자 정보

권역	노인				합계	
	60~69세		70~79세			
	남	여	남	여		
서울/인천/경기	91	90	48	47	139	137
대전/세종/충청/강원	27	27	16	15	43	42
광주/전라/제주	22	22	14	15	36	37
부산/대구/울산/경상	53	54	29	30	82	84
전국	193	193	107	107	300	300

음성 데이터 수집은 별도의 전용 툴을 사용하여 수집하며, 직접 사용자가 즉석에서 듣고 음성 품질을 검토하여 재녹음이 가능하도록 하였다. 또한 문장 단위로 녹음이 진행되어, 음성 데이터 분절 및 검수에 편의성이 향상되도록 하였다.

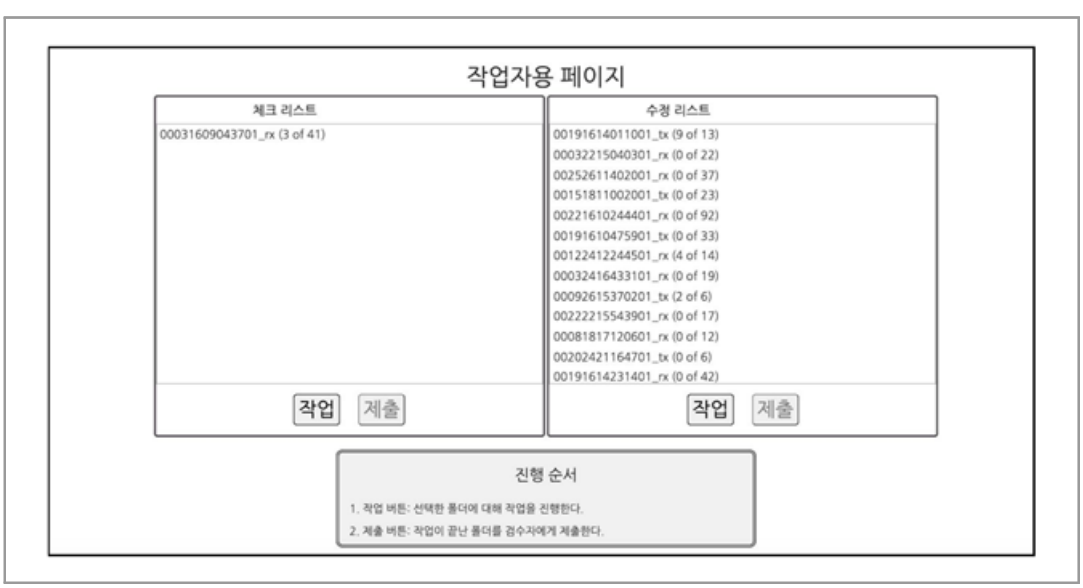


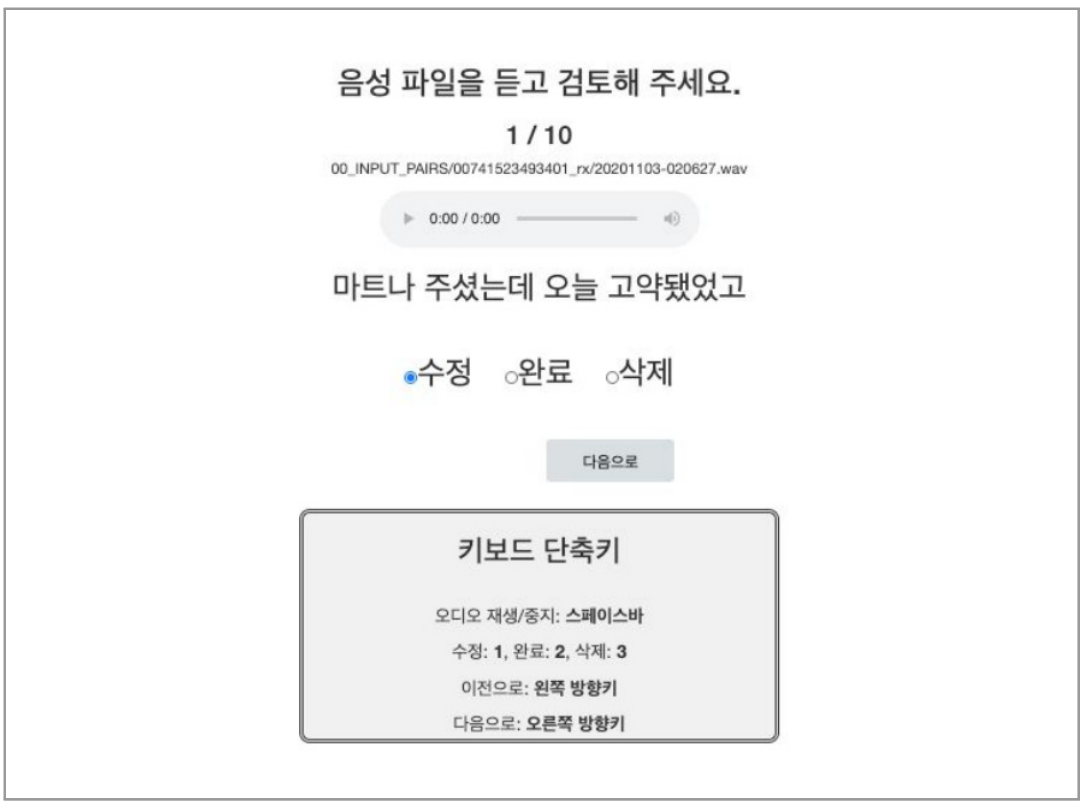


●○ 검수와 품질 확보

대량의 데이터를 높은 품질로 생성하기 위하여, 단가는 낮지만 품질 관리가 어려운 크라우드소싱 방식의 데이터생성 작업을 보완하기 위한 검수 프로세스의 정립은 데이터셋 구축에 매우 중요한 의미를 갖는다. 이 데이터셋에서는 3단계 검수 체계를 구축했는데, 가장 하위 레벨에는 크라우드 워커들이 작업한 결과물을 가이드라인에서 제시한 형식에 맞는지 체크하는 검수자가 있었고, 이들이 검수한 결과물에 대해서 내용적으로 유효한지 검수하는 재검수자가 팀을 이뤄 활동한다.

효율적인 음성 데이터 검수를 위해 별도의 전용 툴을 사용한다.





●○ 데이터 구축 담당자

수행기관(주관) : 미디어젠(주) 송민규 상무님
(전화: 02-6429-7104, 이메일: minks@mediazen.co.kr)