

# 인공지능 데이터 구축·활용 가이드라인

## - 자연어 / 자유대화 주1)-

※ 주1) 지정공모 과제의 경우 소분류 단계인 “세부 데이터명” 건별로 작성 요청, 단 세부 데이터셋이 공통적이고 복수인 경우 부제에 제목 명기 가능.  
(예, 농업 영상 데이터(4)의 경우 4개의 세부데이터명(위성/드론 농경작지 촬영 영상, 농산물 품질 이미지, 시석작물 개체 영상, 주요 농작물 생육 이미지 데이터) 건별로 작성 요청)

인공지능 데이터 구축	사업 총괄	이-이-이
	데이터 설계	이-이-이 Consortium
	원천데이터 수집 및 정제	이-이-이 Consortium
	데이터 가공	이-이-이 Consortium
	데이터 검수	이-이-이 Consortium
	클라우드 소싱	이-이-이 Consortium
	저작도구 개발	이-이-이 Consortium
	AI모델 개발	이-이-이 Consortium
	응용 서비스 개발	이-이-이 Consortium
가이드라인 작성	NHN다이퀘스트	전기왕 책임
	셀바스 AI	윤재선 이사
	아임클라우드	박진석 부장
가이드라인 버전	가이드라인 버전 : v0.4 가이드라인 작성일 : 2020-12-07	

# 목 차

<b>1. 데이터 명세 정보 .....</b>	<b>1</b>
1.1 데이터 정보 요약 .....	1
1.2 데이터 포맷 .....	2
1.3 어노테이션 포맷 .....	3
1.4 데이터 구성 .....	3
1.5 데이터 통계 .....	7
1.6 원시데이터 특성 .....	9
1.7 기타 정보 .....	9
 <b>2. 데이터 구축 가이드 .....</b>	 <b>11</b>
2.1 데이터 구축 개요 .....	11
2.2 문제정의 .....	12
2.3 획득·정제 .....	13
2.4 어노테이션/라벨링 .....	31
2.5 검수 .....	43
2.6 활용 .....	54
 <b>별첨. 데이터 구축 일정 .....</b>	 <b>62</b>

## 1. 데이터 명세 정보

### 1.1 데이터 정보 요약

데이터 이름	자유대화 AI 데이터 - 1. 자유대화 (일반남여) - 2. 자유대화 (노인남여) - 3. 자유대화 (소아남여, 유아 등 혼합) - 4. 한국인 외래어 발화	
활용 분야	- 연구분야 : 음성인식, 음성언어처리, 자연어처리, 한국어 음성언어연구, 신호처리 등 - 산업분야 : 온/오프라인 기반의 음성인식, AI비서, Voice BOT, Voice Command & Control, AI 로봇, 음성인식기반 키오스크	
데이터 요약	1. 자유대화 (일반남여) - 10대에서 50대 사이의 일반인 남녀의 발화 데이터 - 녹음 인원 2,000명 이상, 4,000시간 음성 데이터 2. 자유대화 (노인남여) - 60세 이상의 남녀 발화 데이터 - 녹음 인원 1,000명 이상, 3,000시간 음성 데이터 3. 자유대화 (소아남여, 유아 등 혼합) - 3세~6세, 7~10세 연령의 남녀 발화 데이터 - 녹음 인원 1,000명 이상, 3,000시간 음성 데이터 4. 한국인 외래어 발화 - 녹음 인원 2,000명 이상 4,000시간 음성 데이터 - 한국인이 발성한 외래어가 포함된 음성 데이터	
데이터 출처	- (신규 제작)	
데이터 이력	배포버전	0.4
	개정이력	-
	작성자/ 배포자	전기왕 책임

## 1.2 데이터 포맷

### ○ 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 상세정보 (성별 / 연령 / 지역)</li> <li>녹음환경 정보 (실내 / 실외 : 대중교통, 거리 등)</li> <li>대화 주제 및 상세내용</li> </ul>

- 원천데이터(음성파일)과 메타데이터(Json)로 구분
- 원천데이터(음성파일)은 각각의 파일명으로 구분 (Ex. sample1.wav)

### ○ 메타데이터 형태 (Json)

```
{
  "발화정보" : {
    "stt" : "밥 배 따로, 디저트 배 따로 몰라? ",
    "scriptId" : "일반통합-23615",
    "fileNm" : "일반남여_일반통합15_F_1520872503_37_수도권_온라인_23615.wav",
    "recrdTime" : "3.330",
    "recrdQuality" : "16K",
    "recrdDt" : "2020-11-04 10:33:53",
    "scriptSetNo" : "DQ"
  },
  "대화정보" : {
    "recrdEnvrn" : "실내",
    "colctUnitCode" : "AI 챗봇",
    "cityCode" : "수도권",
    "recrdUnit" : "iOS",
    "convrsThema" : "세계의 디저트"
  },
  "녹음자정보" : {
    "gender" : "여",
    "recorderId" : "1520872503",
    "age" : 37
  }
}
```

### 1.3 어노테이션 포맷

대분류	속성표기	의미	타입	필수여부
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	
대화정보	colctUnitCode	수집방법	String	Y
	convrsThema	대화주제	String	Y
	cityCode	지역	String	Y
	recrdEnvrn	녹음환경	String	Y
	recrdUnit	녹음도구	String	Y
녹음자정보	recorderId	녹음자ID	String	
	gender	성별	String	Y
	age	나이	String	Y

### 1.4 데이터 구성

#### 1.4.1 데이터 구조

- 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)
  - ※ 자유대화의 경우 스크립트없이 자유 주제로 대화하기 때문에 스크립트번호는 'P'로 기재
- 영역구분 : 일반남여 / 노인남여 / 소아남여 / 외래어
- 스크립트셋번호 : 일반통합00 / 노인대화00 / 소아남여00 / 외래어00 / 자유대화(자유주제)
  - ※ 소아남여는 다음과 같이 추가적인 구분을 가져감
    - 3~6세 : 소아남여00/소아남여0000 (셋번호구분 2자리: 1000문장내외/4자리 2000문장 내외)
    - 7~19세 : 소아남여00
  - ※ 자유대화의 경우 스크립트없이 자유 주제로 대화하기 때문에 스크립트셋번호는 'P'로 기재
- 성별 : M(남성) / F(여성)
- 사용자ID : 각 녹음 사용자 식별 값
- 나이 : 숫자(정수)

- 지역 : 수도권 / 충청 / 강원 / 경상 / 전라 / 제주 /기타
- 녹음위치 : 실내 / 실외 / 녹음실
- 스크립트번호 : 숫자(정수)



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

일반남여	일반통합01	M	1456144760	37	전라	실내	012587
노인남여	노인대화03	F	3205411230	71	제주	실내	541357
소아남여	소아남여04	M	8454121576	11	서울	실외	874411

영역 구분	스크립트 세번호	성별	사용자ID	나이	지역	녹음 위치	스크립트 번호
-------	-------------	----	-------	----	----	----------	------------

#### 1.4.2 데이터 분류 (세부 분류 내역은 '별첨' 주제분류표 참조)

## ○ 일반남여

대분류	소분류	건수	주제	건수	주제	예시
건강	39	54	- 간헐적 단식 효과 - 내장비만 줄이는 방법			
경제	3	9	- 집, 살까? 말까? - 미성년 자녀 주식 투자			
교육	7	9	- 암기 잘하는 법 - 대안학교 선택			
기술	1	1	- 유관순 얼굴 복원			
날씨	3	5	- 장마철 습기 제거법 - 긴 장마의 어려움			
동물	2	10	- 생긴 거랑 다르게 잔인한 동물 - 애완견 이름 짓기			
문화	10	12	- 비틀즈 vs 롤링스톤즈 - 코로나 이후 일본에서의 유행			
사회	33	52	- 타인의 의견에 공감하기 - 성차별적인 광고			
쇼핑	10	11	- 유통기한 지난 썬크림 활용법 - 오토바이 구매하기			

여가	15	59	- 돌레길 여행 - 아내와 함께 운동하기
요리	3	8	- 에어프라이어 사용법 - 카레에서 고지 잡내 없애기
음식	26	32	- 줄어드는 쌀 소비량 - 비건식단과 채식식단의 차이
일상	106	193	- 마음의 안식처를 찾아서 - 바닥에 떨어진 음식을 먹어도 될까
직장	19	37	- 센스있는 건배사 - 해외에서 시차 적응 잘하는 방법
취미	17	20	- 넷플릭스 요금제 - 음식 사진을 잘 찍고 싶다면
기타	-	-	- 자유대화(자유주제) 등

## ○ 노인남여

대분류	소분류 건수	주제 건수	주제 예시
TV	4	6	- 드라마 - 영화
날씨	2	2	- 날씨 - 일기예보
쇼핑	4	9	- 한의원 - 노트북
안부·일상대화	8	104	- 가족건강 - 여가생활
전공	1	2	- 영어 - 콩쿠르
정치경제	5	19	- 아파트시세 - 아르바이트
취미	11	42	- 콘서트 - 반려견
기타	-	-	- 자유대화(자유주제) 등

## ○ 소아남여(3~6세)

대분류	소분류 건수	주제 건수	주제 예시
일상	84	1196	- 할아버지 - 인라인스케이트
자연	2	6	- 여름 - 달
기타	-	-	- 자유대화(자유주제) 등

## ○ 소아남여(7~10세)

대분류	소분류 건수	주제 건수	주제 예시
가정	3	14	- 질문 및 요청 - 끝말잇기
명절	2	12	- 선달그믐 - 보름달
몸과 마음	10	23	- 내일이 왔으면 좋겠어요 - 기분이나 감정 표현하기
생일	5	9	- 생일파티 준비 - 원하는 선물
소개	5	5	- 가족 소개하기 - 자기가 좋아하는 것 소개
여행	1	6	- 봉평 - 불산

일상	108	449	- 학교 생활 보고 - 어른이 되고 싶은 이유
장소	1	14	- 학교 주변 - 영화관
특별한 날	11	19	- 놀이공원 - 엄마 생일
하루 일과	11	18	- 아침에 일어났을 때 - 세수와 양치질을 할 때
학교	13	37	- 방학에 한 일 - 존경하는 인물
기타	-	-	- 자유대화(자유주제) 등

## ○ 외래어

대분류	소분류 건수	주제 건수	주제 예시
그리스어	7	15	- 디오니소스(Dionysos) - 스피нк스(Sphinx)
네덜란드어	10	18	- 하이네켄(heineken) - 케빈 더브라위너(Kevin De Bruyne)
덴마크어	2	5	- 프리츠 한센(Fritz Hansen) - 브라네르(Branner, Hans Christian)
독일어	32	141	- 골드만삭스(Goldman Sachs) - 글로켄슈필(Glockenspiel)
라틴어	16	27	- 레보도파(levodopum) - 불레우테리온(buleuterion)
러시아어	12	27	- 닥터 지바고(Doctor Zhivago) - 블라디미르(Vladimir)
산스크리트어	9	14	- 브라마(Brahma) - 브라마굽타(Brahmagupta)
스웨덴어	5	8	- 구스타브 아돌프(Gustav Adolf) - 북스테후데(Buxtehude, Dietrich Diderik)
스페인어	26	65	- 라 벤타나(La Ventana) - 삼보앙가(Zamboanga)
아랍어	11	21	- 칼루아 커피(Kahlua Coffee) - 데네볼라(Denebola)
영어	384	7,204	- 마이에스큐엘(MySQL) - 드렁큰 타이거(Drunken Tiger)
이탈리아어	23	122	- 산타 마리아(Santa Maria) - 델리카토(delicato)
인도어	3	6	- 고푸람(Gopuram) - 란치(Ranchi)
일본어	49	169	- 도라에몽(ドラえもん) - 고시엔(甲子園)
중국어	17	66	- 궈바로우(鍋包肉) - 란창강(瀾滄江)
페르시아어	4	5	- 바자회(bazar會) - 니샤푸르(Nishapur)
포르투갈어	5	13	- 사르가소해(Sargasso海) - 펠레(Pele)
프랑스어	49	227	- 베베미뇽(BeBe Mignon) - 에뛰드하우스(Étude House)
한국어	3	5	- 이태원 클라쓰(Itaewon class) - 종이컵(종이cup)
한자	35	59	- 장애물(障礙物) - 안내 데스크(案内 desk)
히브리어	5	8	- 아멘(amen) - 할렐루야(hallelujah)
기타	19	64	- 고클로비치(Gombrowicz, Witold) - 마추픽추(Machu Picchu)



## 1.5 데이터 통계

### 1.5.1 데이터 구축 규모

과제명	주요 내용	데이터 수집 방법	데이터 구축량	데이터 형식
데이터1 자유대화(일반남여)	10~50대의 화자의 음성 데이터 (남녀비율 1:1)	오프라인(스튜디오) 온라인 (음성채팅)	2000명 이상의 화자 4000시간	음성데이터/ 텍스트데이터 (음성과 매칭)/ 관련정보
데이터2 자유대화(노인남여)	60세 이상의 노인 화자의 음성 데이터 (남녀비율 1:1)	오프라인 (스튜디오/인터뷰) 온라인 (스마트스피커)	1000명 이상의 화자 3000시간	음성데이터/ 텍스트데이터 (음성과 매칭)/ 관련정보
데이터3 자유대화(소아남여, 유아)	3~6세, 7~10세의 소아/유아 화자의 음성 데이터 (남녀비율 1:1)	오프라인(스튜디오) 온라인(온라인 녹음)	1000명 이상의 화자 3000시간	음성데이터/ 텍스트데이터 (음성과 매칭)/ 관련정보
데이터4 한국인 외래어 발화	화자의 외래어 발화 음성 데이터	오프라인(스튜디오) 온라인(온라인 녹음)	2000명 이상의 화자 4000시간	음성데이터/ 텍스트데이터 (음성과 매칭)/ 관련정보

### 1.5.2 데이터 분포

- 녹음자 성별 및 연령대 인원 및 비율

- 데이터1 (일반남여)

성별 인원(비율)	
남자	여자
1	1

- 데이터2 (노인남여)

성별 인원(비율)	
남자	여자
1	1

- 데이터3 (소아남여)

성별 인원(비율)		연령대 인원(비율)	
남자	여자	3~6세	7~10세
1	1	2	8

- 데이터4 (외래어)

성별 인원(비율)	
남자	여자
1	1

## 1.6 원시데이터 특성

### 1.6.1 대상분류

- 실제 : 실제 사람의 발성으로 녹음된 음성 데이터

### 1.6.2 제약조건

- 일부 제약있음 : 일부 음성 데이터 수집 시 사전 스크립트 있음

### 1.6.3 속성

- 음성 데이터 : PCM (WAV)
- 메타데이터 : Json

## 1.7 기타정보

### 1.7.1 포괄성

- 성별 : 남 / 여 (1:1 비율)
- 연령 : 10대 미만 ~ 60대 이상
  - 일반 : 10대 / 20대 / 30대 / 40대 / 50 대
  - 노인 : 60대 이상
  - 소아 : 3~6세 / 7세~10세
  - 외래어 : 10대 / 20대 / 30대 / 40대 / 50대 / 60대 이상
- 지역 : 수도권, 충청, 강원,

### 1.7.2 독립성

- 원시데이터에 의존 사항 없음

### 1.7.3 유의사항

#### ○ 저작권 이슈

- 녹음에 사용된 시나리오(스크립트)는 자체제작으로 음성파일 배포시에 별도의 저작권 이슈 없음

#### ○ DB 구축시 예상되는 문제점 해결 방안

- 노인과 소아의 경우 일반 성인에 비해 구축시간이 2~3배 걸리는 등의 문제가 발생할 소지가 있으므로 노인남여와 소아 데이터 구축시 유의사항에 따라 구축 작업 필요
- 노인과 소아의 경우 각 1,000명 이상의 화자를 확보해야 하므로, 작업량에 따른 보상 비율을 일반 보다 높게 책정하여 작업 동기 부여
- 기존 일반 작업자의 노인과 소아 가족 참여를 독려하여, 노인, 소아 작업자 모집 및 기존 일반 작업자를 통한 교육 진행
- 노인 및 소아 작업자는 소규모 모집이 어려운 특성이 있어, 인력이 확보된 기관(노인대학, 보육원 등)과 연계하여 각 기관에 소속된 인력을 단체로 확보

#### ○ 노인남여 데이터 구축

- 노인의 경우 발화 앞 뒤로 묵음과 간투어가 빈번한 특성이 있음
- 음성 데이터 녹음 시 스크립트 문장을 길게 하여 묵음과 간투어를 최소화 하여 녹음 작업 수행

- 1명의 작업자가 2시간의 작업량을 채우기 어려우므로 전체 녹음시간을 2시간에서 1시간 30분으로 줄여서 작업 결과물의 완성도를 높이고 보다 많은 노인 일자리를 창출

○ 소아남여 데이터 구축

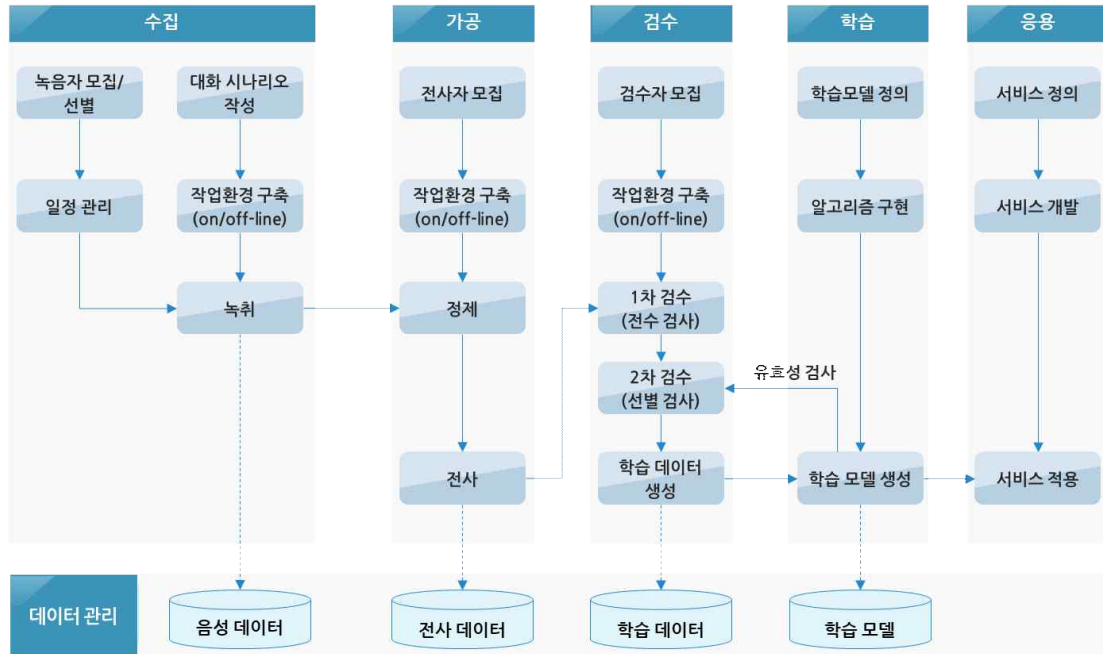
- 소아의 경우 긴 스크립트 문장을 제공할 경우, 발화시 문장 후반부에 발음이 엉키는 등의 문제가 발생할 수 있으며 집중을 이어가기 어려울 수 있음
- 짧은 스크립트를 제공하여 짧은 문장을 발화하여 녹음 수행
- 1명의 작업자가 2시간의 작업량을 채우기 어려우므로 전체 녹음시간을 2시간에서 1시간 30분으로 줄여서 작업 결과물의 완성도를 높이고 보다 많은 소아 일자리를 창출
- 소아작업자가 구축한 데이터를 정제 후 획득된 데이터가 일반, 노인에 비해 구축량이 적을 것으로 판단되므로, 보다 많은 소아 작업자를 추가 확보하여 화자 수 증강

1.7.4 관련 연구

-

## 2. 데이터 구축 가이드

### 2.1 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 적응 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

## 2.2 문제정의

### 2.2.1 임무 정의

- 인공지능(AI) 기반 한국어 음성인식 서비스 활성화를 위한 자유대화(일상대화) 지식 데이터 구축
- 실제로 사용하는 방대한 분량의 자유대화를 효과적으로 인식하기 위해 인공지능(AI) 기반 한국어 자유대화(일상대화) 데이터를 구축하며, 국민들에게 더욱 질 높은 인공지능(AI) 서비스를 제공 할 수 있는 양질의 학습데이터 확보하여 기술적 기반을 마련
- 영아/어린이, 노인층의 발화특성을 반영한 자유대화(일상대화)데이터를 구축하여, 음성인식 기반 인공지능(AI) 서비스 사용에 대하여 소외되는 계층이 없는 인프라 구성
  - 일반 성인남녀의 대화데이터를 사용할 경우 노인, 영아, 어린이 등의 발화의 특성을 고려되지 않기에 음성인식 기반 서비스가 정상적으로 제공되지 않는 가능성이 존재하며, 음성인식 기반 인공지능 서비스 사용에 대한 소외 계층이 발생 할 수 있음

### 2.2.2 데이터 구축 유의사항

- 데이터 구축 시 외부 데이터 활용 하는 경우 법적/사회적 이슈가 없도록 데이터 무상 확대 제공에 대한 라이선스를 득하여 확약 처리
- 데이터 제공 업체 또는 기관의 대표 명의 공문 혹은 확약서로 명문화
- 데이터 지원 확약서 작성 시 양식에 따라 데이터명, 데이터 제공 기간, 데이터 활용 과제명 등의 정보를 명시

## 2.3 획득·정제

### 2.3.1 원시데이터 선정

#### ○ 원시데이터 수집을 위한 대화시나리오 선정

- 작성 필요성: 발화 수집목적에 부합하는 대화주제 및 원고개발을 통해 음성수집 목표 달성
- 작성절차: 수집기준 선정-> 수집-> 정제 -> 분류 -> 발화원고 및 주제작성

작성절차	 일반남녀	 노인남녀	 소아남녀, 유아 등 혼합	 한국인 외래어 발화
 수집기준 선정	수집항목 선정 · 일상 대화 주제 분류 · 대화 상황 및 장소, 공간 선정	수집항목 선정 · 일상 대화 주제 분류 · 대화 상황 및 장소, 공간 선정	수집항목 선정 · 3~5세 누리과정 생활주제, MCDI-K 기반 선정 · 국립국어원 <초등학생을 위한 표준 한국어 의사소통> 참고	수집항목 선정 · 외래어 단어 및 해당 단어를 사용한 용례 발화 수집 · 콘텍스트와 관련한 인명, 지명, 제목 포함 수집
 수집	생활 어휘 수집 · 인터넷, 유튜브, SNS 검색 · 웹 크롤링을 통한 수집 · 한국어 교재분석	노인 어휘 수집 · 장년층 대상 방송 프로그램 대본 분석 · 웹 크롤링을 통한 수집 · 방언 관련 논문 분석	소아 어휘 수집 · EBS 어린이 방송 대본 분석 · 소아대상 유튜브 검색 · 초등학교 국어교과서, 동화책 주제, 소재, 성취기준 분석	외래어 어휘 수집 · 인터넷, 유튜브, SNS, 블로그를 검색 · 웹 크롤링을 통한 수집
 정제	생활어휘 정제 · 생활어휘 사용빈도 분석 · 민감한 이슈 발원 제외	노인 어휘 정제 · 노인 어휘 사용빈도 분석 · 대상연령 적합 어휘 선별	소아 어휘 정제 · 소아 어휘 사용빈도 분석 · 대상연령 적합 어휘 선별	외래어 정제 · 외래어 사용빈도 분석 · 인명, 지명 최소화
 분류	생활어휘 분류 · 대화주제별 카테고리 분류 · 비문법적 표현, 신조어, 말줄임, 반복 등 대화 선정	노인 어휘 분류 · 대화상황별 분류 · 비문법적인 표현, 말 줄임, 노인 특유의 부정확한 발음 어휘 선정	소아 어휘 분류 · 대화상황별 분류 · 비문법적인 표현, 말 줄임, 소아 특유의 부정확한 발음 어휘 선정	외래어 분류 · 유래어 분류 · 주제별 카테고리
 발화원고 주제 작성	대화 시나리오 작성 · 성별, 연령별 주제 선정 · Small Talk 대화, Voice 챗봇 등 상황별, 장소 별 작성	대화 시나리오 작성 · 억양 및 단어 사투리 반영 · Small Talk 대화, Voice 챗봇 등 상황별, 장소 별 작성	대화 시나리오 작성 · 3~6세, 7~10세 주제 선정 · 낭독 발화, 주제별 대화, 자유발화, Small Talk 대화 등 원고 작성	대화 시나리오 작성 · 유래어 분류 · 주제별 카테고리 · 성별, 연령별 어휘

- 일반남녀 발화목록 작성방안
  - 일반 남녀 언어 특징을 고려하여 대화 주제, 상황 별 대화 시나리오 작성
  - 대화 시나리오를 보고 읽는 낭독 발화에 대한 대화 시나리오 작성
  - 자유발화 : 제시된 주제를 바탕으로 다수의 사용자가 자유롭게 채팅을 하는 형태로 대화를 진행  
자연스럽게 대화를 할 수 있도록 지인 또는 가족간 대화 방을 구성할 수 있도록 대  
화방 내 초대 기능 활용
  - 발화분량 : 총 2시간의 발화분량으로 진행하나, 발화자의 노인 발화자의 건강 및 집중력을 고려  
하여 녹음 시간 단위를 1시간 단위로 제한하거나 한 번에 수집할 수 있는 분량을 제한하는 것
- 노인남녀 발화목록 작성 방안
  - 노인 연령대의 언어 특징을 고려하여 대화 주제, 상황 별 대화 시나리오 작성 필요
  - 대화 시나리오를 보고 읽는 낭독 발화에 대한 대화 시나리오 작성
  - 발화분량 : 총 2시간의 발화분량으로 진행하나, 발화자의 노인 발화자의 건강 및 집중력을 고려  
하여 녹음 시간 단위를 1시간 단위로 제한하거나 한 번에 수집할 수 있는 분량을 제한하는 것  
을 고려

발화 종류	발화분량	발화방법	발화 내용
낭독 발화	1시간 1200단어/800문장	다양한 일상 대화를 2사람 이상이 대화 화는 대화체 주제 별로 나누어 역할 별로 읽게 하거나, 단문의 내용을 읽게 함	문장 낭독: 한국어의 음운이 고루 실현되도록 작성한 문장 - 각 대화 스크립트 별 PBS(Phonetically Balanced Sentences) 문장을 추가 문단 낭독: 정보문이나 스토리 등의 내용으로 작성한 문장
자유 발화	1시간	주제만 제시하고 자연스럽게 자발적으로 대화를 자유롭게 끌어가도록 함	자유 발화: 제시된 주제에 대해서 자유롭게 대화 발화 ex) 가족 소개, 아끼는 물건, 애완동물 등

## 발화목록 작성 방안(일반/노인 남녀)

- 소아남녀 발화목록 작성 방안
  - 연령대의 인지 발달, 언어 수준을 고려하여 대화 주제, 상황 별 대화 시나리오 작성 필요
  - 대화 시나리오를 보고 읽는 낭독 발화에 대한 대화 시나리오 작성
  - 소아의 경우 문장의 비율을 50%이상으로 녹음 지문을 확보
  - 발화분량 : 총 2시간의 발화분량으로 진행하나, 발화자의 집중력과 녹음 상태를 고려하여 녹음 시간 단위를 1시간 단위로 제한하거나 한 번에 수집할 수 있는 분량을 제한하는 것을 고려

대상	대화 주제	발화목록 예시
3~6세	<ul style="list-style-type: none"> <li>▪ 교육부 3~5세 누리과정의 생활주제를 중심으로 대화 주제 및 내용 선정</li> <li>▪ 한국의 영유아가 사용하는 어휘를 설계한 MCDI-K(MacArthur Communicative Development Inventory-Korean)를 참고하여 단어를 1차적으로 선정하고 대화 주제 반영</li> </ul>	<b>&lt;낭독 발화&gt;</b> <ul style="list-style-type: none"> <li>▪ 단어(그림 보고 말하기)</li> <li>▪ 문장(듣고 따라하기, 질문과 대답)</li> <li>▪ 주제별 발화 예시) 좋아하는 것 말하기, 색깔 말하기 등</li> <li>▪ 영상을 보고 말해 보기, 챗봇과 대화하기 등</li> </ul>
7~10세	<ul style="list-style-type: none"> <li>▪ 국립국어원 기획 [초등학생을 위한 표준한국어 의사소통] 교재를 중심으로 대화 주제 선정</li> <li>▪ 교육과정 국어 초등학교 1-2학년군 및 3-4학년군 성취 기준 및 국어교과서 참고</li> </ul>	<b>&lt;낭독 발화&gt;</b> <ul style="list-style-type: none"> <li>▪ 문장(초등 교과서나 동화책)</li> <li>▪ 문단(초등 교과서나 동화책)</li> <li>▪ 주제별 발화 예시) 자기소개, 물건 사기, 여행 등</li> </ul>

## 발화목록 작성 방안(소아남녀)

- 한국인 외래어 발화목록 작성방안

- 한글이나 국어로 대체되어 쓰이는 경우가 더 많은 어휘 제외 (인위적으로 정한 순화어는 예외)
- 비속어, 은어 등 제외
- 아직 일반화되지 않은 외국어의 한글 발음이나 표기 제외
- 발화분량 : 총 2시간의 발화분량으로 진행하나, 발화자의 집중력과 녹음 상태를 고려하여 녹음 시간 단위를 30분 단위로 제한하거나 한 번에 수집할 수 있는 분량을 제한하는 것을 고려
- 외래어 단어 및 해당 단어를 사용한 문장단위의 용례발화 포함하여 수집 (모든 문장이 영어로된 문장은 20% 비율 이하로 제한)
- 콘텐츠(영화, 노래)와 관련된 인명, 지명, 제목 포함

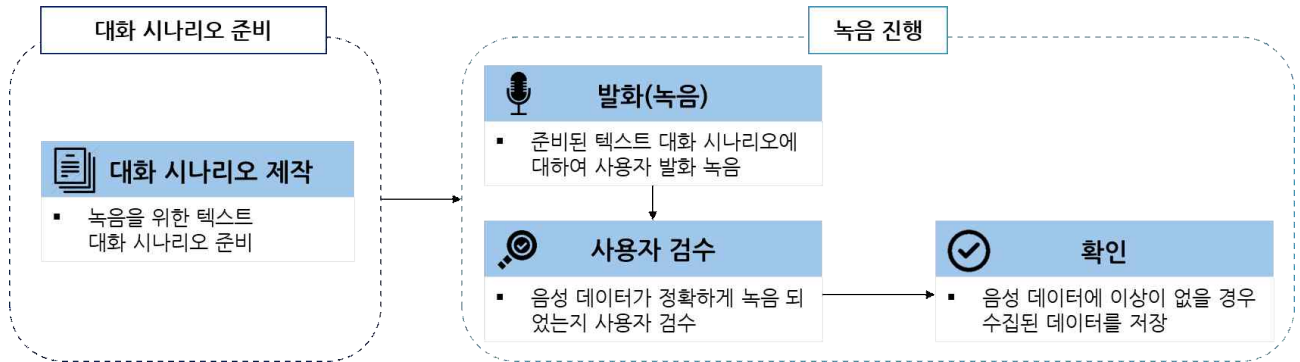
발화 종류	세부 내용
외래어 어휘 수집	<ul style="list-style-type: none"> <li>인터넷, SNS, 블로그를 검색하여 최대한 많은 외래어 수집</li> <li>웹 크롤링/클리닝을 통한 수집 (Python, Perl 등의 언어로 텍스트 프로세싱 대화 시나리오를 생성하여 진행)</li> </ul>
외래어 목록 정제	<ul style="list-style-type: none"> <li>외래어 사용빈도 분석: 가능한 한 대용량의 그리고 복수의 말뭉치(코퍼스)를 이용하여 각 외래어 마다 사용빈도를 분석</li> <li>사용빈도를 정렬(sorting)하여 일정 값(threshold)을 기준으로 제한</li> <li>컷오프 기준은 데이터의 양, 수집 시간 등을 고려</li> </ul>
외래어 유형 태깅 (categorization and tagging)	<ul style="list-style-type: none"> <li>유형 설정: 인명, 지명, 상품명, 프로그램 명, 전문 용어, 스포츠 용어 등</li> <li>유래어 확인: 그 외래어의 유래 언어 검색: 영어, 불어, 일본어 등</li> <li>유형과 유래어에 관한 태깅: 준자동 태깅을 진행하거나, 수작업으로 진행</li> </ul>

#### 발화목록 작성 방안(외래어)

- 대화 시나리오 작성 시 고려사항
  - 시나리오 작성 시 음성 인식에 적합하도록 Phone balance를 고려하여 작성
  - 대화 시나리오에 따라 음성 데이터의 정합성 및 다양성에 영향을 끼치므로, 과제 시작 후 대화 시나리오 작성 내용 검토
  - 일정 길이 이상의 문장으로 이루어져 있는지, 단순 단어 형식인지 여부 검토 (발화의 길이가 밸런싱 있게 녹음이 될수 있도록 고려하여 대화 시나리오를 구성)
  - 시나리오 작성 건이 다른 콘텐츠(소설, 영화 및 드라마 대본)의 내용을 그대로 차용했는지 여부 등을 검토하여 저작권 위반 여부 별도 검증하고, 시나리오 작성 시 타 기관의 콘텐츠가 필요한 경우 협의를 통해 데이터 제공 약약을 작성하고, 해당 콘텐츠 제공 받아 작성
  - 시나리오 작성 과정에서 실제 녹음작업을 수행할 업체의 실무자를 통해 음성인식의 적합한 형태인지를 검증
  - 동일한 내용을 녹음하여 중복되는 데이터가 다수 발생 않도록 충분한 시나리오를 확보하며 시나리오가 부족하지 않도록 녹음이 진행되는 과정에서도 추가적으로 시나리오를 작성될 수 있도록 확인



### 2.3.2 획득·정제 절차



- 발화 : 녹음하기 버튼을 통해 준비된 대화 시나리오를 사용자가 발화
- 사용자 검수 : 녹음된 음성을 들어보고 수정이 필요한 경우 다시 녹음
- 확인 : 녹음된 음성데이터가 이상이 없는 경우 저장

### 2.3.3 획득·정제 기준

- 녹음 수집 데이터 적/부 판별 기준
  - 샘플링 주파수 16kHz
  - 16Bit Resolution 기준으로 오차  $\pm 30,000$  이상의 샘플이 20% 이상인 데이터는 불량으로 판정
  - 발화와 발화 사이의 무음 구간이 2초 이상인 데이터는 불량으로 판정
  - 발화문장 앞뒤의 불필요한 무음은 제외 (단 발화 앞뒤로 100~200msec 정도의 묵음은 음성 인식을 위해 필요)
- 데이터 검수 과정에서 수집데이터의 비율 중 10% 정도의 데이터 불량이 발생할 것으로 예측 (1차 검증 : 전수검사 / 2차검증 : 선별검사)
- 데이터 가공 및 정제 과정에서 음원의 왜곡이 없도록 진행
- 불량 데이터가 발생할 것과 추가 수요가 발생할 것을 감안하여 고려하여 원시데이터(녹취데이터) 수집 과정에서 구축데이터 목표 분량의 5~10%를 추가 수집

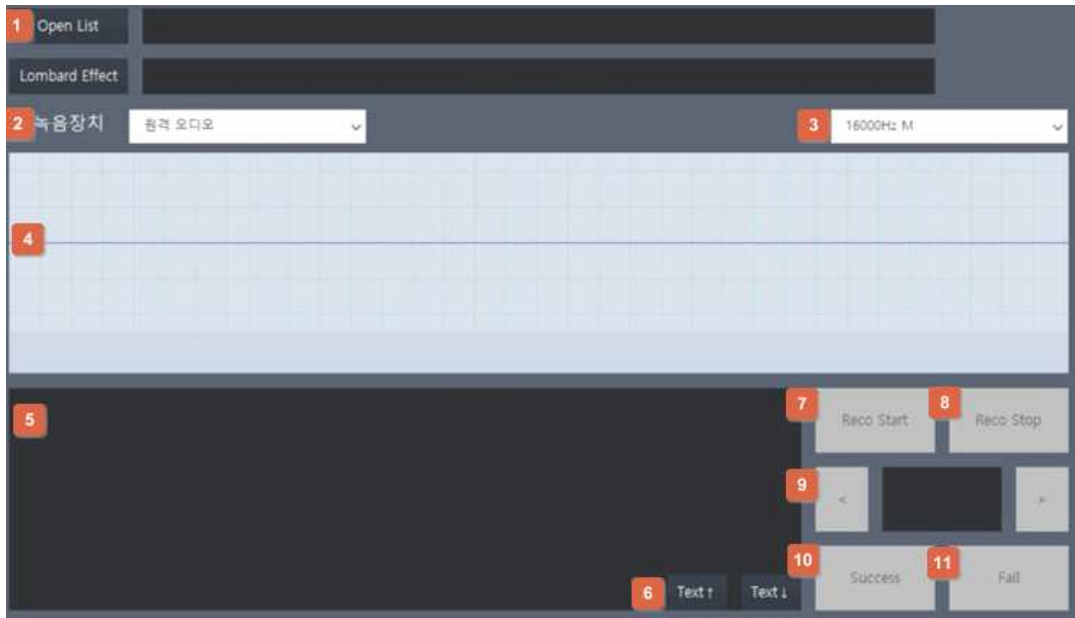
### 2.3.4 획득·정제 조직 *(필요 시 작성)*

-

### 2.3.5 획득·정제 도구

#### ○ PC 녹음도구

- Windows PC 환경에서 녹음 작업
  - 녹음실에 지정된 발화 스크립트를 보고 발성
  - 관리 매니저가 녹음자가 오발화한 경우, 재녹음을 진행하여 발화오류 없는 녹음 데이터 확보 가능
- 녹음 도구 버튼 설명



- ① 녹음 Script OPEN( \*\*\*\*.txt)
- ② 녹음 Device 선택
- ③ 녹음 샘플링 주파수 선택
- ④ 녹음되는 파형 Display
- ⑤ 발화 Script Display
- ⑥ Script font size 조정
- ⑦ 녹음 시작
- ⑧ 녹음 중간 멈춤
- ⑨ Script 앞 뒤 변경
- ⑩ 녹음자가 정상 발화 하면 Success 다음 발화 이동
- ⑪ 녹음자가 오발화하면 Fail 동일 Script 재발화

• 녹음 절차

1) 대본 파일 제작

- 대본 스크립트를 ASCII 타입으로 아래와 같은 포맷의 txt파일을 제작
- 다섯자리 숫자열Wt(TAB)한글 script

```
00000  느라나 줄연
00001  건물을 부수다
00002  철학적인 문제
00003  주먹을 쥐다
00004  물건들을 치우다
00005  도박에 빠지다
00006  왜 그런 줄 알아요
00007  유권자의 선택
```

2) 화자 설정

- 연령대 / 성별 / 이름 등록한다.
- 추후 필요 정보로 변경 예정

화자 설정

연령대 : 성인

성별 : 남자

유창성 : 일반

이름 :

저장

### 3) 녹음 준비 상태

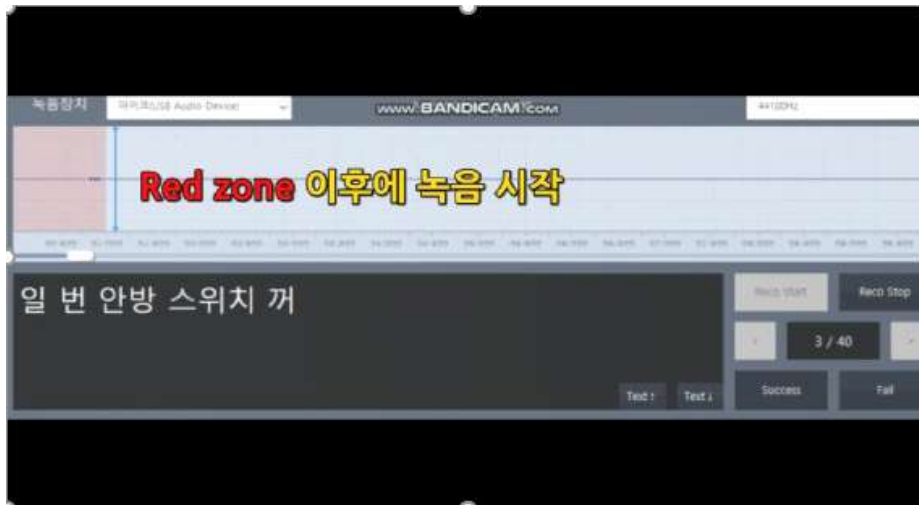
- 발화 리스트 첫 번째 scrip가 화면에 표시한다.
- 총 발화 수 중 현재 정보 표시한다.
- Rec Start 버튼만 활성화된다.



- ① 첫번째 발화 Script
- ② 녹음 활성화
- ③ 현재 녹음 위치 정보

### 4) 녹음 Process

- Red zone 이후 음성 발화 및 녹음

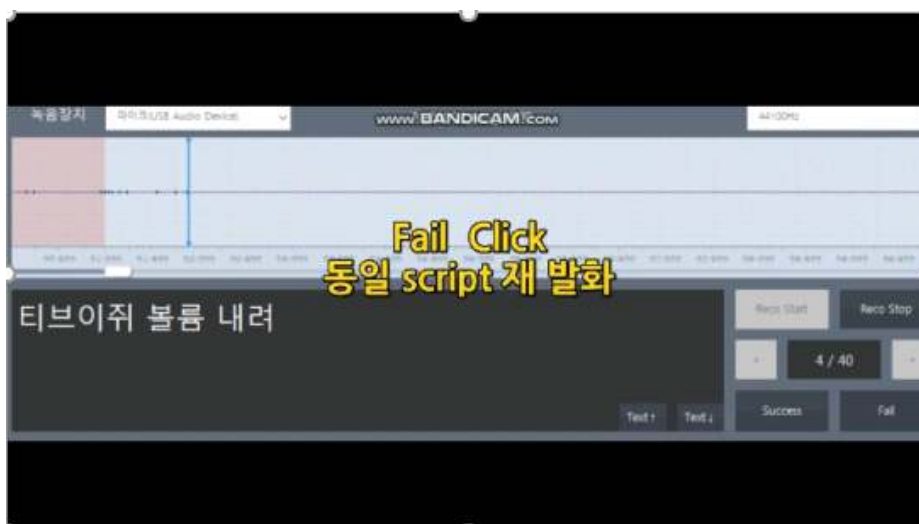


- 음성 발화 끝난 후, 1 ~ 2 초 녹음 멈춤



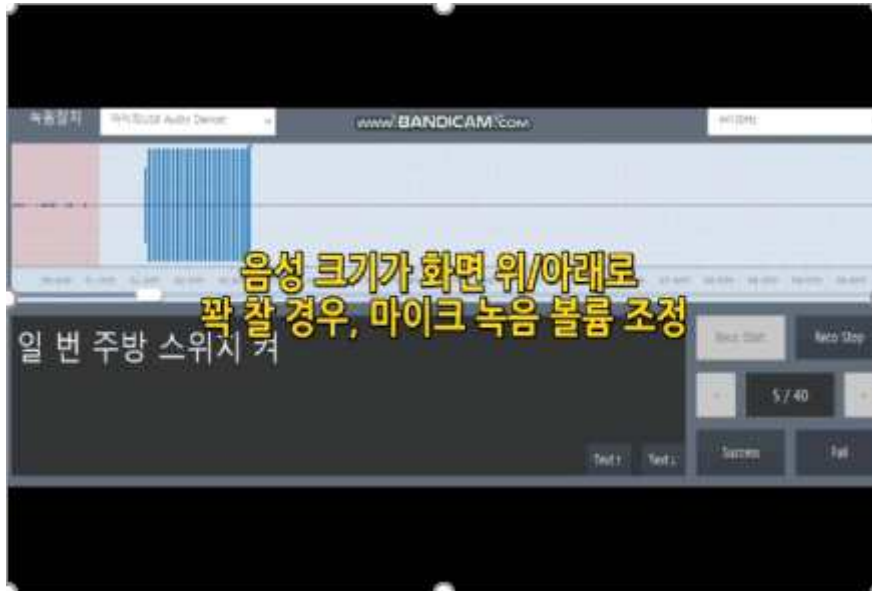
##### 5) 오발화 음성 후 Process

- Fail Click 후, 재 녹음



## 6) 마이크 볼륨 조정

- Wave Display 위 아래로 꺾어서 녹음이 되는 경우, 마이크 녹음 볼륨 조정



- 활성화된 녹음 DEVICE 선택



- 수준 속성에서 마이크 볼륨 조정



## • 클리핑 방지 등의 녹음 가이드 정책

- 녹음실에서 음성 DB 수집 시, PC 녹음 매뉴얼의 녹음 가이드 정책을 따라 작업 수행
- 대화 앞뒤의 Silence Margin 없이 발성을 방지하기 위해 음성 녹음 시작 지점에 0.5초 구간의 Red Zone을 두어, 이 후에 발성하도록 시각적으로 표시

- Waveform 형식으로 시각적으로 클리핑이 될 정도로 모든 음성을 크게 발성하는 경우는 재녹음을 진행
- 근래의 Android, iOS 단말의 코덱 성능으로는 32,000 이상으로 크게 녹음되는 경우, 찢어지는 소리 없이 AGC를 이용하여 음성인식으로 입력으로 사용 가능
- 클리핑된 음성이 녹음되는 경우, 음성인식 학습 측면을 고려하여 다양한 조건의 음성 DB 확보가 강인한 인식 성능 보장 가능

○ 모바일 녹음도구1 (클라우드웍스)

• 녹음 작업자

1) 모바일에 클라우드웍스 어플을 설치 후 문장 녹음 프로젝트를 클릭하여 녹음



- 프로젝트 참여를 위해서는 반드시 클라우드웍스에 가입 필요
- 가입한 ID 공유 및 프로젝트 참여 그룹으로 설정
- 녹음용 핸드폰에 클라우드웍스 어플을 설치 후, 작업 시 마다 로그인

2) 프로젝트 카드를 클릭하여, 작업을 진행한다.

<
외래어 문장 음성 수집

1

작업 완료 : 10 건

문의하기


문장을 보고, 녹음해주세요.

2

기말고사 대신 갤러리에 다녀와 감상문을 제출하면 됩니다.

버튼을 누르고 녹음을 진행해주세요.

3



자연스럽게 발화해주세요

4

저장

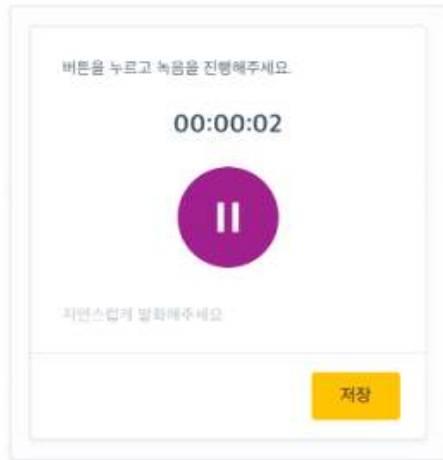
이력 보기

5

작업 제출

- ① 작업완료 : 작업 제출을 눌러 완료한 작업 건 수 표시
- ② 문장 /스크립트 : 미리 소스 세팅 된 문장 / 스크립트 노출 영역. 해당 영역을 보고 발화함
- ③ 녹음 : 버튼을 눌러 녹음 활성화. 문장을 모두 발화한 후 일시정지 버튼 클릭

### < 버튼을 누른 후 화면 >



### < 일시정지 누른 후 화면 >



- ④ 저장 : 녹음된 데이터 저장
- ⑤ 작업 제출 : 저장 완료된 데이터 최종 제출

#### • 녹음 검수자

- 1) 크라우드웍스 웹사이트에 로그인하여 녹음된 데이터를 직접 검수
  - 검수를 진행할 담당자의 크라우드웍스 아이디를 공유
  - 검수자 권한 및 검수 그룹 설정
  - 본인의 작업 데이터는 본인이 검수 불가능(다른 작업자의 데이터만 검수 가능)
  - 반드시 크롬 브라우저로 접속

crowdworks

1,367P 유지예 남 PREMIUM [→] LOGOUT

검수 리스트

최신순

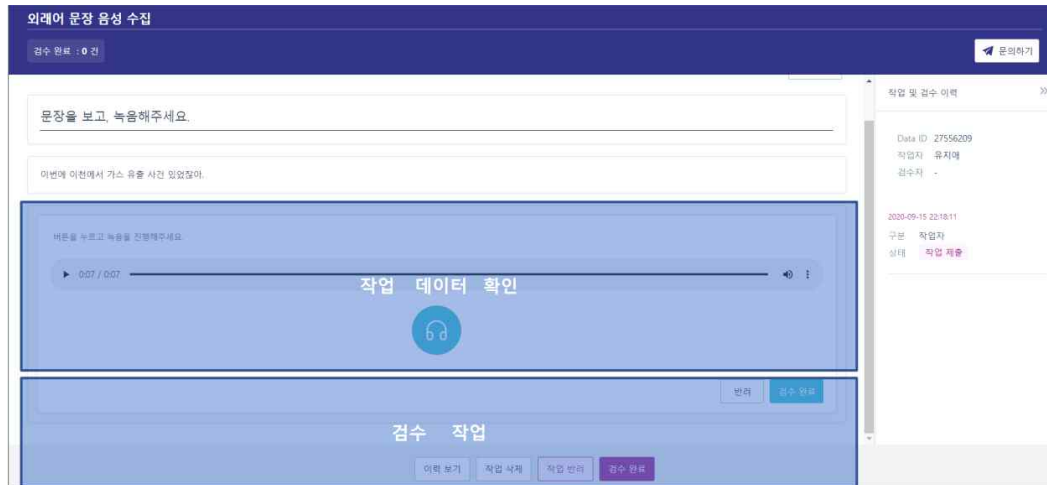
검수 작업으로 최고의 수익을 창출하는 방법이 궁금하세요? 검수 필수 지침 →

외래어

번호	프로젝트명	작업명	시작일	작업/전체	반려	검수완료	재검수대기	검수대기	검수진행	작업자	검수자
8864	외래어 문장 음성 수집	외래어 문장 음성 수집	2020.09.15	10 / 100	0	0	0	10	0	1	0

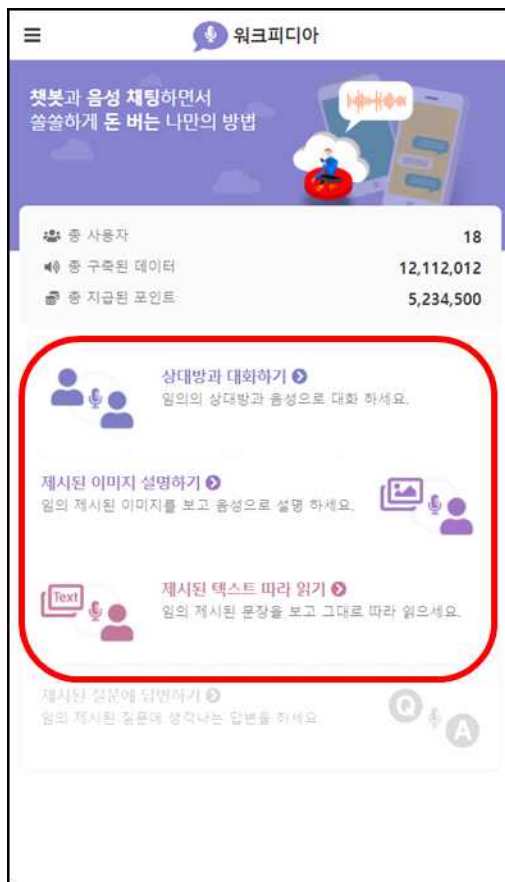
- 2) 작업명 클릭 시 검수자에게 검수 데이터 할당
- 3) 녹음된 데이터를 듣고 가이드에 맞지 않을 경우, 반려버튼을 클릭
  - 반려 사유는 작업자가 이해할 수 있도록 상세하게 작성
- 4) 반려 사유를 적은 뒤 하단 작업 반려를 클릭





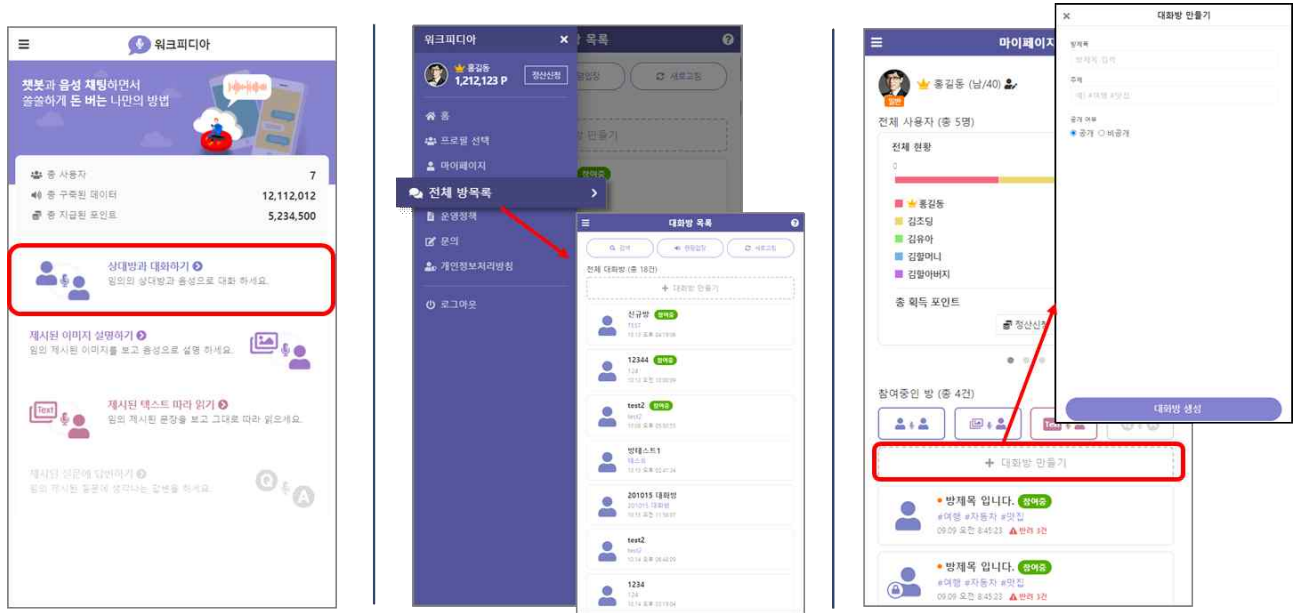
○ 모바일 녹음도구2 (NHN다이렉스트)

• 작업항목 선택



- 홈 메뉴에서 상대방과 대화하기/제시된 이미지 실행하기/제시된 텍스트 따라 읽기 중 원하는 작업 항목을 선택
- 마이페이지에서 참여중인 방의 작업 항목을 선택
- 상대방과 대화하기
- 홈화면에서 상대방과 대화하기 선택하여 대화방에 참여

- 좌측 메뉴의 전체 방목록에서 원하는 대화를 선택하여 대화방에 참여
- 마이페이지에서 새로운 대화방을 생성하여 대화방에 참여

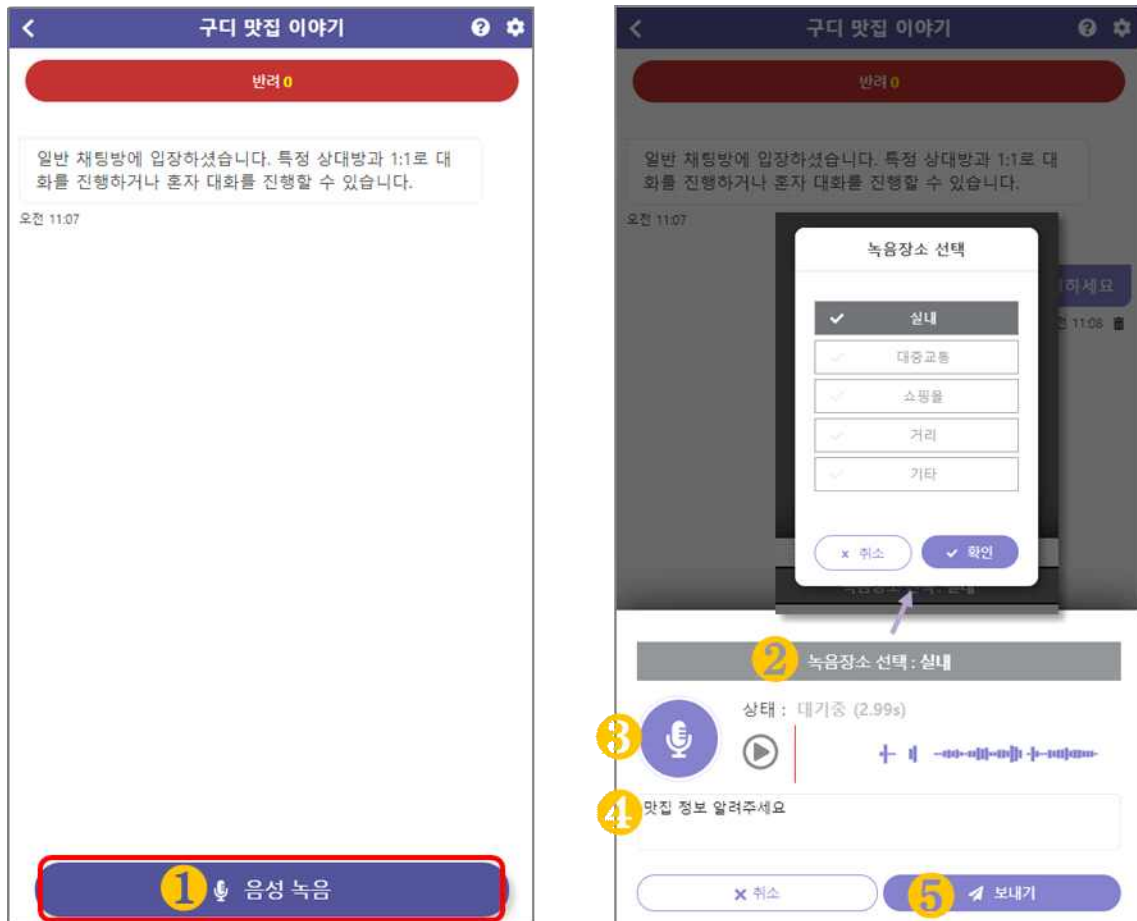


<홈: 상대방과 대화하기>

<좌측메뉴: 원하는 대화방 선택>

<마이페이지: 대화방 생성>

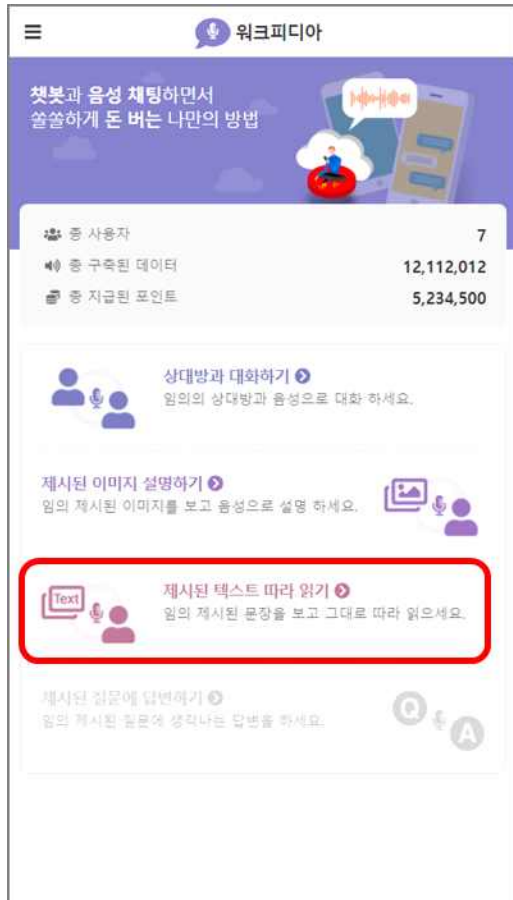
- 대화를 통한 녹음 방법



① 음성 녹음 버튼을 눌러 대화 시작

- ② 녹음 장소 선택
- ③ 녹음 아이콘 버튼을 눌러 녹음 시작
- ④ 하단에 생성된 텍스트 확인 후 오류 수정
- ⑤ '보내기' 버튼 눌러 전송

- 제시된 텍스트 따라 읽기
  - 홈화면의 제시된 '텍스트 따라 읽기' 버튼 눌러 참여
  - 마이페이지의 '텍스트 따라 읽기' 아이콘 누른 후 하단의 참여하기 버튼 눌러 참여

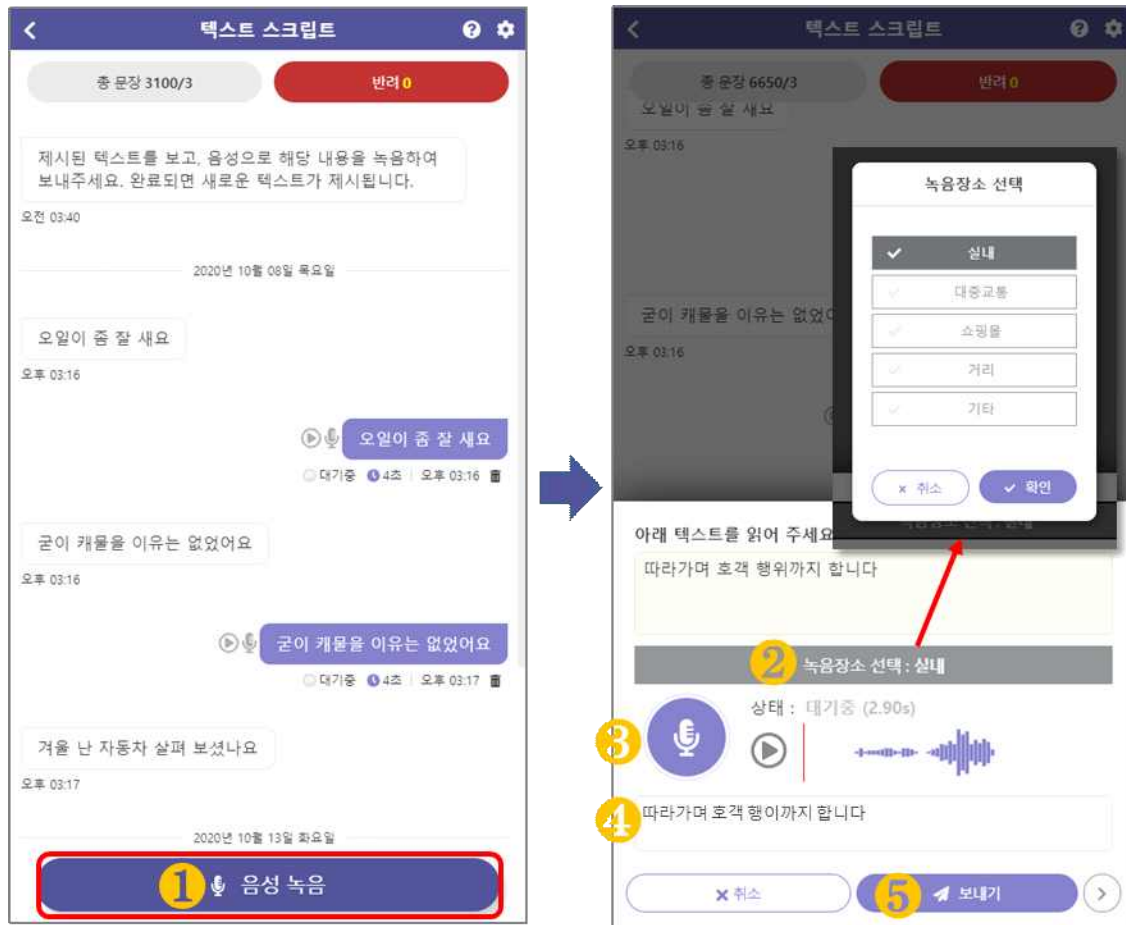


<홈: 제시된 텍스트 따라 읽기>



<마이페이지: 텍스트 따라읽기>

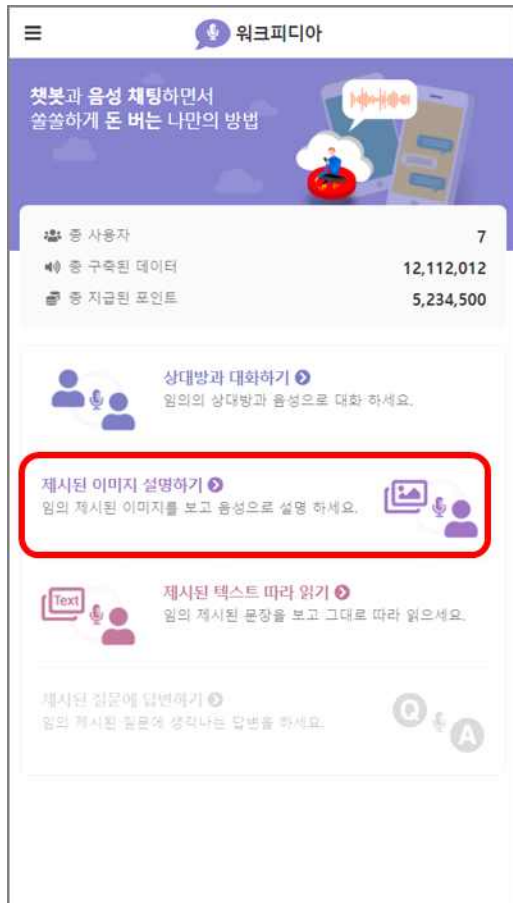
- 제시된 텍스트 따라 읽기를 통한 녹음 방법



- ① 음성 녹음 버튼을 눌러 대화 시작
- ② 녹음 장소 선택
- ③ 녹음 아이콘 버튼을 눌러 녹음 시작
- ④ 하단에 생성된 텍스트 확인 후 오류 수정
- ⑤ '보내기' 버튼 눌러 전송

• 제시된 이미지 설명하기

- 홈페이지의 제시된 '제시된 이미지 설명하기' 버튼 눌러 참여
- 마이페이지의 '제시된 이미지 설명하기' 아이콘 누른 후 하단의 참여하기 버튼 눌러 참여

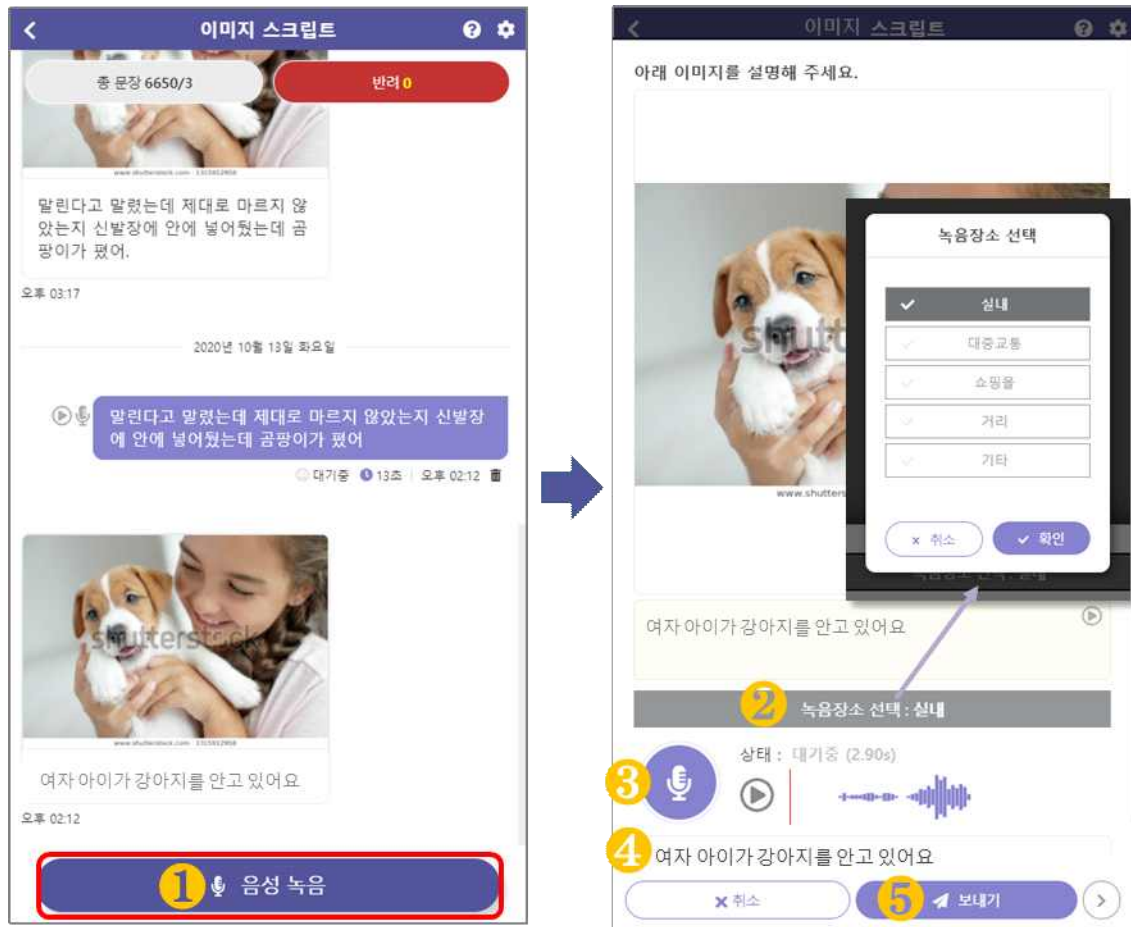


<홈: 제시된 이미지 설명하기>



<마이페이지: 제시된 이미지 설명하기>

- 제시된 이미지 설명하기를 통한 녹음 방법

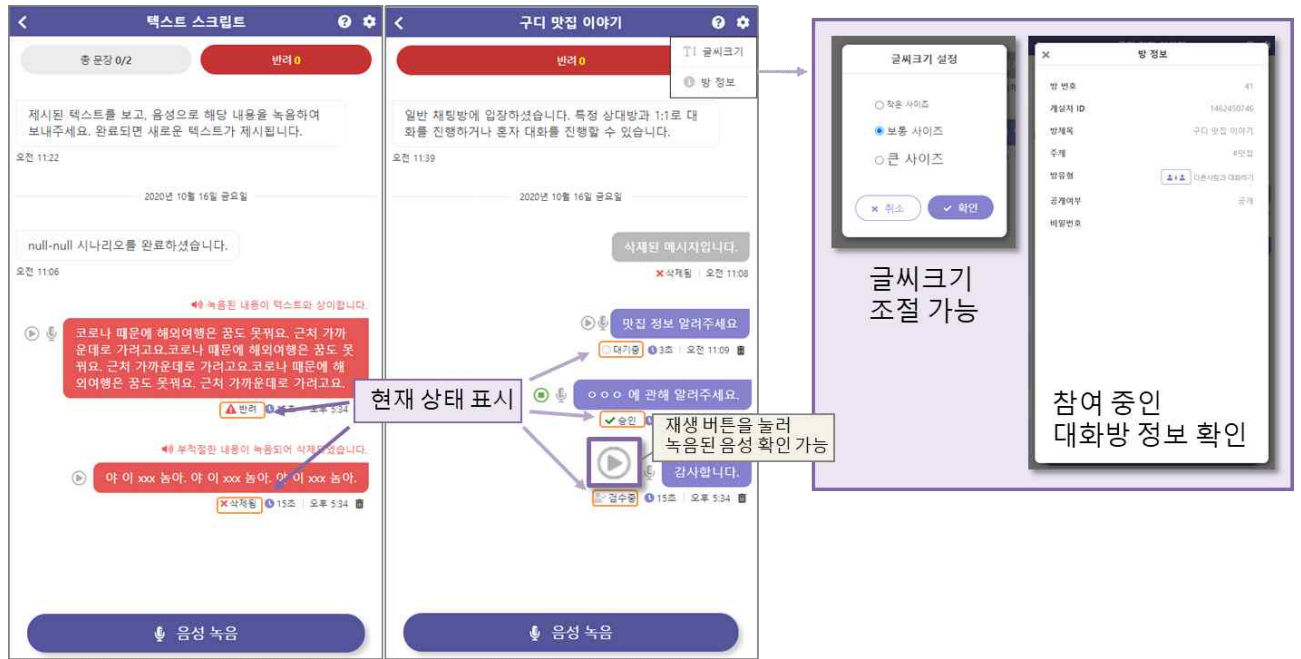


- ① 음성 녹음 버튼을 눌러 대화 시작
- ② 녹음 장소 선택
- ③ 녹음 아이콘 버튼을 눌러 녹음 시작
- ④ 하단에 생성된 텍스트 확인 후 오류 수정
- ⑤ '보내기' 버튼 눌러 전송

• 주의사항

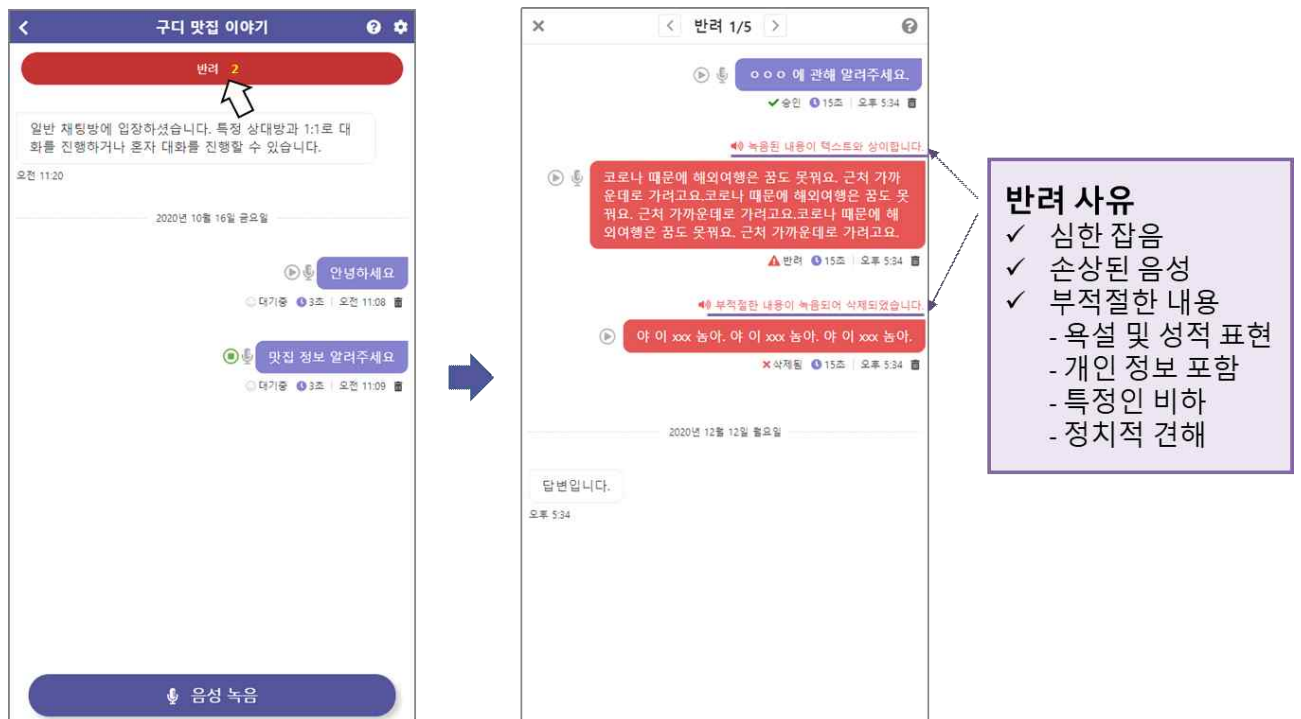
- 심한 잡음 및 불분명한 음성은 재녹음 요청 가능
- 정치적 견해, 개인정보 포함, 특정인 비하 및 비속어, 성적표현 등의 부적절한 내용과 재녹음이 이루어지지 않은 반려데이터는 삭제 가능

• 대화방 정보



- 대화방 상단: 총 문장/반려 건수 표시
- 말풍선 하단 표시 정보
  - 현재 진행 상태(대기중/검수중/승인/반려)
  - 녹음시간
  - 발화 삭제 버튼: 검수전(대기중) 상태인 경우만 가능

#### • 반려



- 반려 건수가 표기된 버튼을 눌러반려 목록 확인
- 반려 항목 재녹음시 재검수→ 검수 후 승인 시 포인트 지급



## 2.4 어노테이션/라벨링

### 2.4.1 어노테이션/라벨링 절차

#### ○ 음성인식(전사) 작업 절차

- ① 입력되는 음성 데이터의 샘플링 주파수 설정
  - 16kHz
- ② 녹음 데이터의 지정
  - 전사를 위한 음성 파일과 발화 텍스트 파일을 지정
- ③ 전사자는 수정 edit 창에 음성을 듣고, 녹음자의 발성이 정확하게 철자전사 되어 있는지 확인 후, 틀린 발성이 있는 경우는 텍스트를 수정
- ④ 음성인식 학습을 위해 사용되는 데이터이기 때문에, 음성 구간의 앞과 뒤에 1초정도의 무음 구간이 존재해야 하나, 지정된 시간보다 짧은 무음 구간이 있을 경우에는 데이터의 무음 구간을 복사하여 붙여 넣을 수 있는 기능 제공
- ⑤ Studio에서 녹음DB 이외의 실 환경에서 수집 받은 음성데이터에는 외부 잡음 등이 포함되어 있으니, 외부 잡음 Filler Noise를 철자 전사 정보에 추가
  - (NO:) 내일 오후 세시에 예약해 주세요
- ⑥ 자유 발화인 경우, 녹음자도 모르는 사이에 발성하는 어, '음', '글세' 등도 음성인식 학습에 필요한 정보이기 때문에 간투어 Filler Pause 정보를 철자 전사 정보에 추가
  - (FP:음) 내일 오후 세시에 예약해 주세요
- ⑦ 유아, 노인들이 발성할 경우, 전사자가 들었을 때 명확하게 철자전사 작업을 하지 못할 음성 전사는 발성오류 정보를 추가
  - 내일 오후 (SP:세시에) 예약해 주세요
  - : 세시, 네시 가 명확하게 들리지 않는 경우
- ⑧ 마이크로폰 근접에서 발성하는 경우, 녹음자의 들숨, 날숨, 웃음 소리 등이 녹음되는 음원의 경우는 화자 잡음 Speaker Noise를 철자 전사 정보에 추가
  - (SN:) 내일 오후 세시에 예약해 주세요 (SN:)
- ⑨ 모든 입력 버튼, Play command는 마우스 움직임을 통한 click 실행보다는 시간 단축을 위해 단축키를 설정하여 전사 작업을 수행
  - Play : CTRL + Space
  - 화자잡음 : SHIFT + F1
- ⑩ 일반적으로 정의된 전사 규칙 외에 정보 입력이 필요할 경우, MEMO 창에 입력
- ⑪ 파일별로 전사 작업이 완료되면, 음성데이터 파일명과 동일한 TRS 파일을 생성



```
[Text Information]
The Original EPD Start=0
The Original EPD End=440
The Original Text= 내일 오후 세시에 예약해 주세요
The Modified EPD Start=0
The Modified EPD End=440
The Modified Text= (FP:음) 내일 오후 세시에 예약해 주세요
Memo=
```

- ⑫ 전사자와 검수자는 동일한 전사틀을 사용하며, 검수자가 검수할 경우에는 TRS 파일을 열어 수정된 철자 정보를 열고, 수정 작업을 진행
- ⑬ 스튜디오 녹음 수집 외 온라인 등 다른 방법으로 수집되는 정보를 각 참여기관과 협의하여 진행

## 2.4.2 어노테이션/라벨링 기준

○ 본 사업 산출물인 전사규칙서를 준용하여 수행

- 개요
  - 숫자, 영어, 기호를 사용하지 않고 한글로만 전사
  - 전사 시 전사규칙과 관련된 기호 이외는 비사용
    - 입력 가능한 기호 : (SP:), (FP:), (SN:), (NO:)
  - 표준발성에서 벗어나거나 같은 전사에 대하여 두 가지 이상 발음이 가능한 경우 발음전사 표기
    - 철자전사 : 표준어법에 맞게 표기하고, 음성인식의 언어모델링 등을 주된 목적으로 함
    - 발음전사 : 발성된 내용을 소리 값에 최대한 가깝게 표기하고, 음성인식의 음향모델링을 주된 목적으로 함
  - 단어의 앞과 뒤에 거의 붙어 발생한 잡음은 단어와 분리하여 표기
  - 잡음이 있는 상황에서 사람에게서 발생하는 잡음(입술소리, 숨소리)은 명확히 구분될 정도로 큰 것만 표기
    - 화자 잡음, (SN:) : 웃음소리, 숨소리, 입술소리
    - 외부 잡음, (NO:) : 녹음자 이외의 주변 잡음, 음악소리
  - 띄어쓰기는 한글 맞춤법에 맞도록 하되, 표준어법으로 명확히 결정할 수 없는 경우에 띄움
- 숫자표현
  - 기본적으로 숫자는 모두 숫자 기호가 아닌 한글로 표현
  - 한국어의 경우 십진 단위로 띄어쓰기
  - 숫자를 하나씩 발음한 경우에 띄어쓰기
  - 단위를 나타내는 '년', '월', '일', '시', '분' 등은 붙여쓰기
    - 예) - 오대 그룹이 모여/자동차 다섯대를
      - 이십 사시간(24시간)/스물 네시간(24시간)
      - 팔 육 공에 이 사 삼 칠(860-2437)

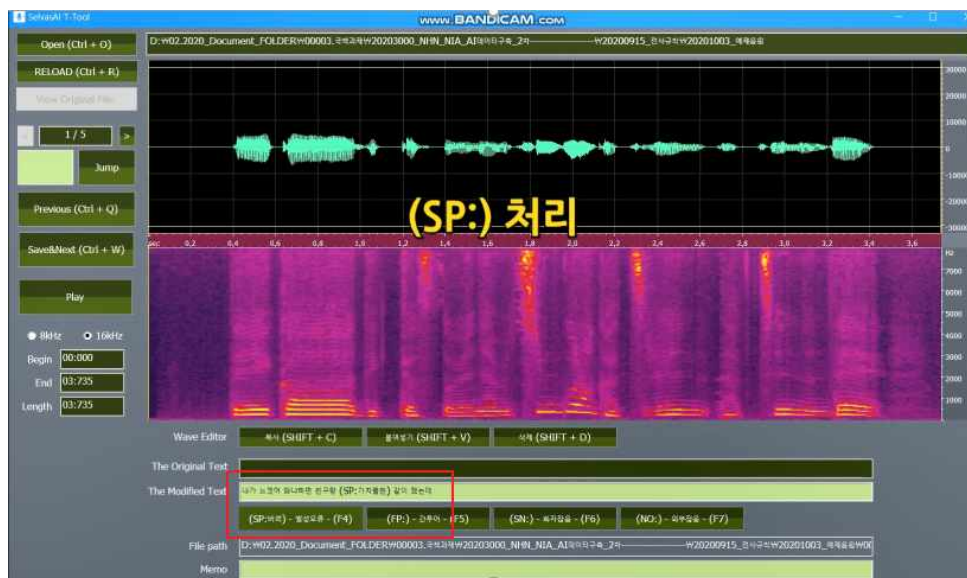
- 십 사시(14시)/열 네시(14시)
- 천 구백 구십 구년에 (1999년에)
- 숫자만으로 이루어진 기념일 등 특정 의미가 있는 단어들을 숫자 단위로 띄어쓰기  
예) - 팔 일 오 (8.15) , 사 일 구 (4.19)
- 오 칠 오 공 부대(5750부대)
- 간투어표현
  - 발성자가 다음 발성을 준비하기 위해서 소요되는 시간을 벌기 위해서 발성하는 것으로 의미 없음
  - 간투어를 포함하여 (FP:\*\*) 로 표기  
예) (FP:에)/(FP:아)/(FP:그)/(FP:어)/(FP:음)/(FP:저)/(FP:저기)/(FP:으) /(FP:응)
- 약어/외래어 표현
  - 약어의 형태의 알파벳을 발화하는 경우, 붙여쓰기  
(하단 알파벳 한글 표기 참조)  
예) 케이비에스 (KBS)/에이티엔티 (AT&T)
  - 된소리 나는 외래어는 표준어 로 철자전사로 표기  
예) 스타벅스(스타벅쓰), 서비스(써비쓰), 센터(쎄터)
  - 우리말로 표기하여 자연스러운 것은 통상적인 한글 표현으로 표기  
예) - 뉴욕, 시카고, 파티
  - 버스, 핸드폰, 모바일, 인터넷, 호텔

알파벳	한글 표기	알파벳	한글 표기
A	에이	N	엔
B	비	O	오
C	씨	P	피
D	디	Q	큐
E	이	R	알
F	에프	S	에스
G	지	T	티
H	에이치	U	유
I	아이	V	브이 / 비
J	제이	W	더블유
K	케이	X	엑스
L	엘	Y	와이
M	엠	Z	지/제트

<알파벳 한글 표기>

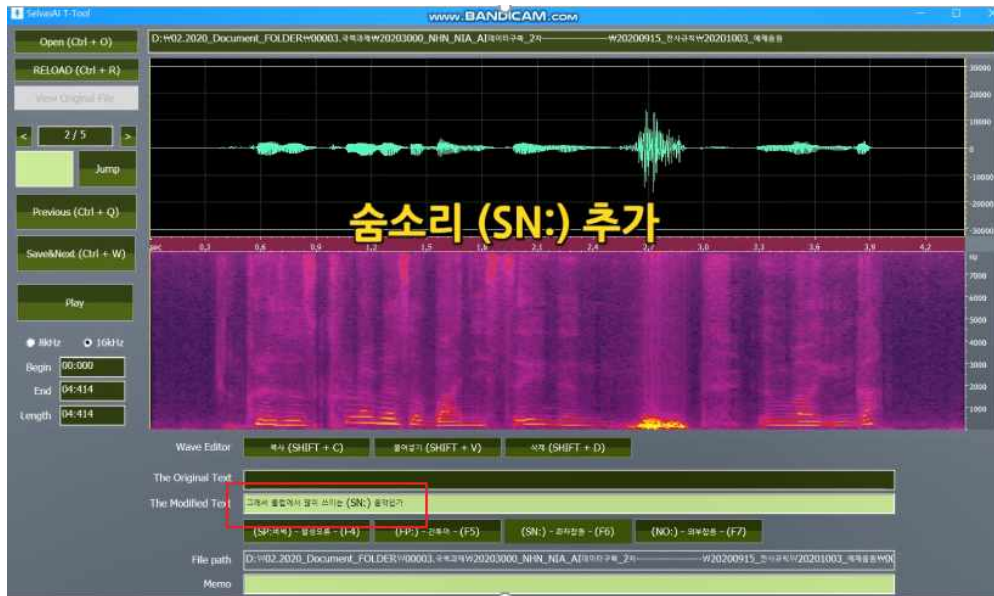
- 비표준발음 표현
  - 어미 비표준 발음으로 들리는 경우 표준어 형태의 철자전사로 표기  
예) - 같아요 ← 같아여, 같애요
  - 했고 ← 했구

- 했고요 < 했구여, 했고여, 했구요
- 입니다 < 입니더
- 밥을 먹었고요 < 밥을 먹었구요, 밥을 먹었구여, 밥을 먹었고여
- 기타
  - 축약 발성은 표준어 형태의 철자전사로 표기  
예) - 안녕하세요 정선희입니다 < 안녕하세요 정선희니다
  - 알아 듣기 힘든 발음, 발성과 동시에 발생된 잡음 처리
    - 화자가 발음한 내용을 잘 알아 듣기 힘들 때, (SP:\*\*)로 표기  
예) 나는 (FP:이럴꼬) 그것을 해결하였다
  - 발성과 동시에 발생하는 외부 잡음은 (NO:\*\*)로 표기  
예) 기차 타는 (NO:곳이) 어디입니까 ('곳이'발성할 때 외부 잡음이 크게 섞임)
  - 반복 발성이나 잘못된 발성은 (SP:\*\*)로 표기  
예) 아침에 (SP:학교) 학교에 갔다
  - 방언에 해당하는 발성은 발음 전사로 표기  
예) 핵교 ( 학교의 방언 )
  - 대화체 문장은 문장 자체가 이상하더라도 발음 전사
  - 버벅 거림 (SP:) 표기



<SelvasAI T-Tool 사용예시 - 버벅 거림 (SP:) 표기>

- 화자 잡음 (SN:) 표기



<SelvasAI T-Tool 사용예시 - 화자 잡음 (SN:) 표기>

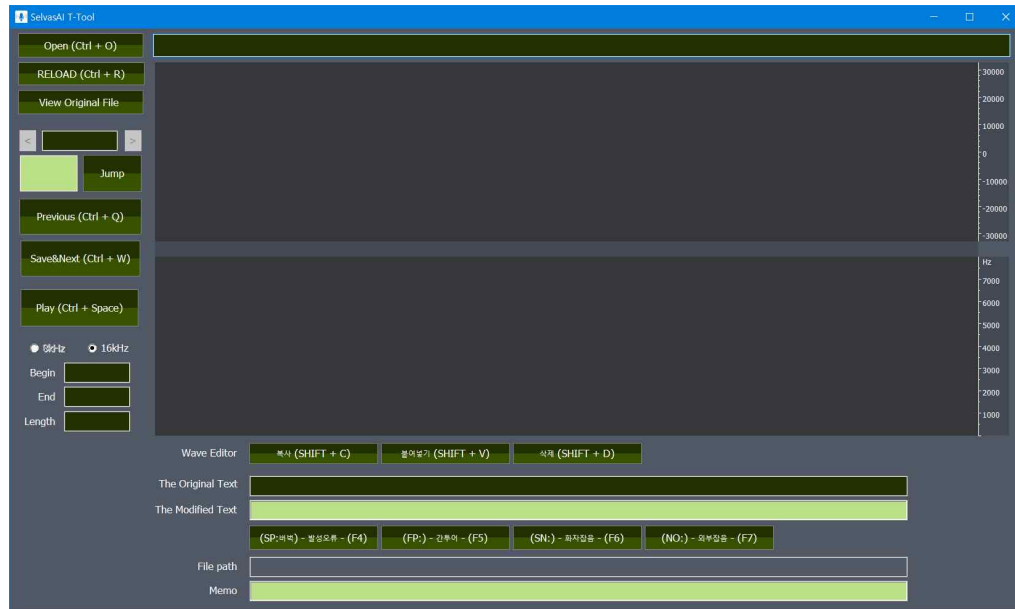
### 2.4.3 어노테이션/라벨링 조직(필요 시 작성)

..

### 2.4.4 어노테이션/라벨링 도구

#### ○ 어노테이션/라벨링 도구1 (셀바스 AI)

- 개요
  - 녹음실에서 지정된 발화 script를 보고 발성
  - 데이터 정제 작업 : SelvasAI T-Tool을 이용하여 녹음자가 발화한 음원의 script를 확인, 수정하는 작업
  - silence 음성 구간 삽입, 삭제, Script 수정, 외부잡음, 화자잡음, 간투어 정보 추가



&lt;SelvasAI T-Tool 화면&gt;

- 사전작업
  - DEVICE44Kto16K : \*.PCM 음원 폴더
  - TEXT : 대본 script
  - 대본 script text를 DEVICE44Kto16K 폴더에 복사



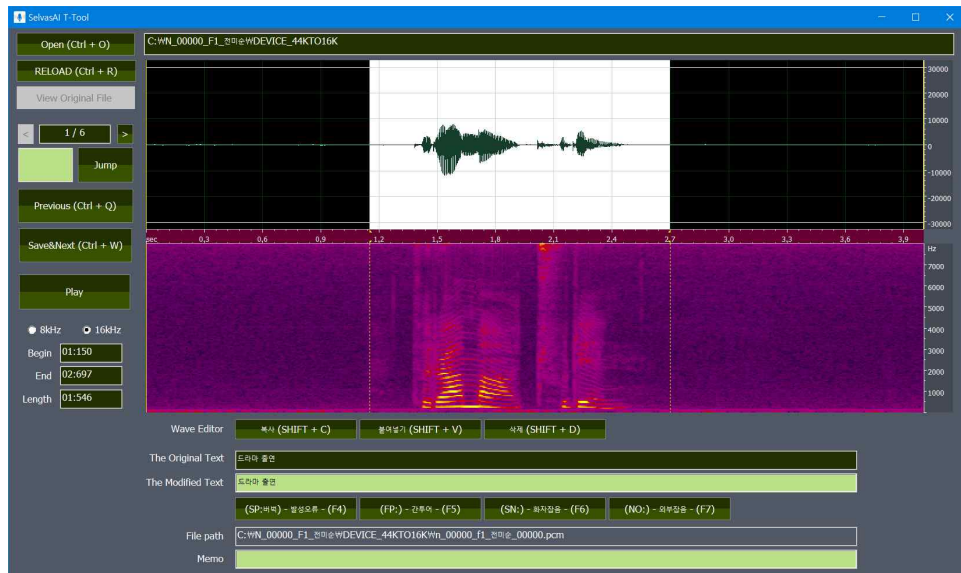
&lt;사전작업&gt;

- 전사도구(SelvasAI T-Tool) 사용법
  - OPEN 버튼을 click 하여 전사할 음원, scrip가 있는 폴더를 선택



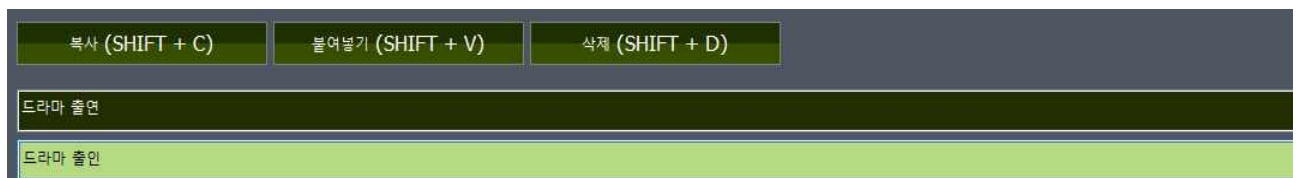
<음원, script 폴더 선택>

- 음성 구간 듣기: 시간 단축을 위해 음성 구간 좌우 0.5초를 포함하여 선택
  - Waveform 영역에 왼쪽 마우스 click 상태에서 drag → 음성 구간 white region으로 반전
  - Play Button click or Left Ctrl + Space



<음원 재생>

- 전사작업
  - The Modified Text 영역에 script 수정
  - 드라마 출연 이라고 발성 ? "드라마 출연" 으로 script 변경



<전사 작업>

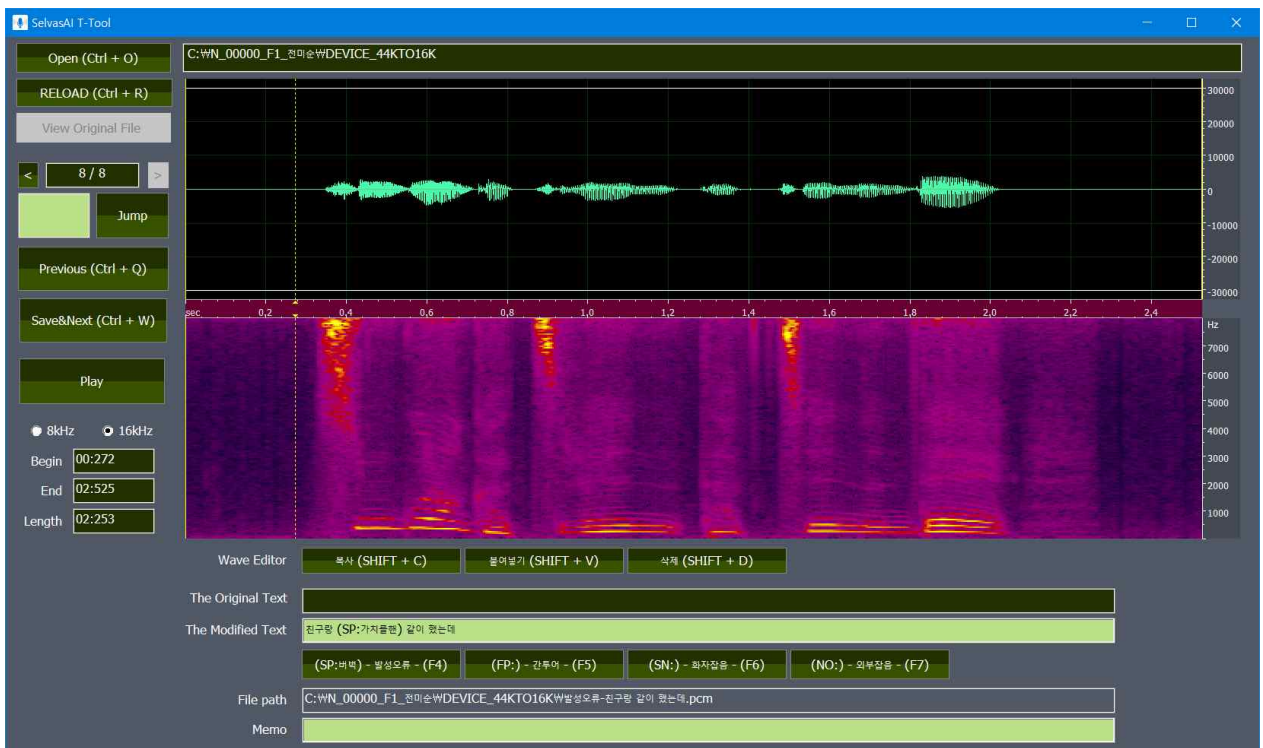


- 3.4. 작업 내용 저장 및 다음 음성 파일 OPEN
  - Save&Next button click or CTRL+W



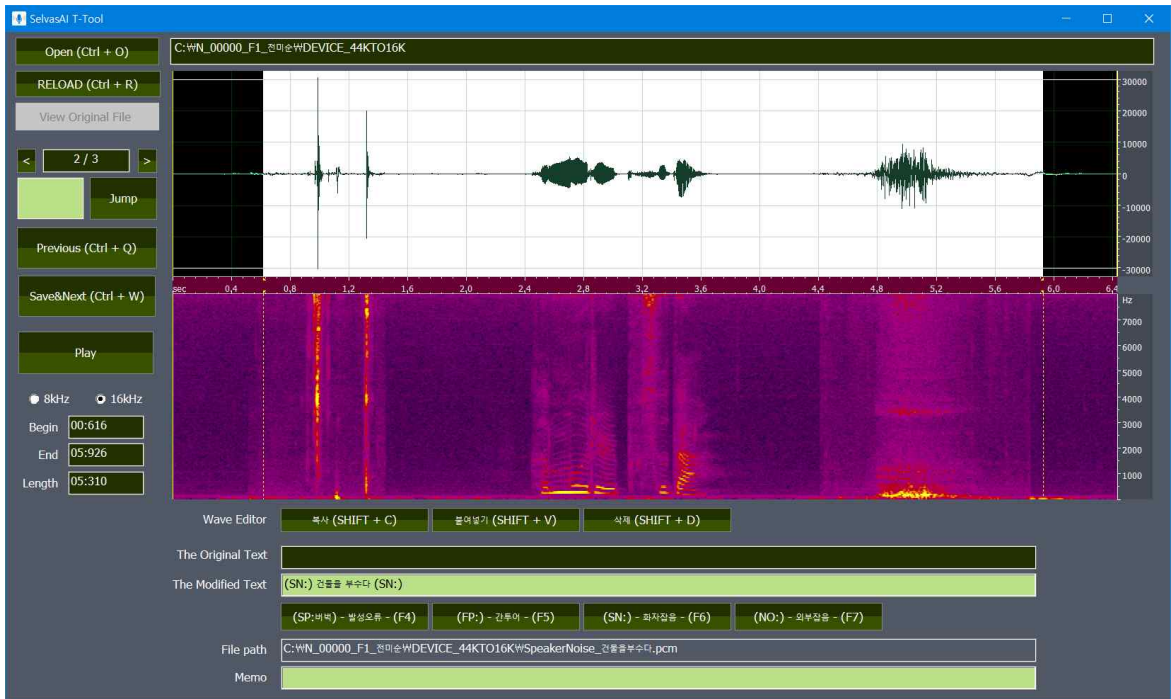
<저장 및 다음 음성 파일 OPEN>

- 잡음, Filter 정보 추가 방법
  - 발성 오류, 버벅 거리는 음원은 소리나는대로 전사하고 SP Filler 추가



<잡음 및 Filter정보 추가>

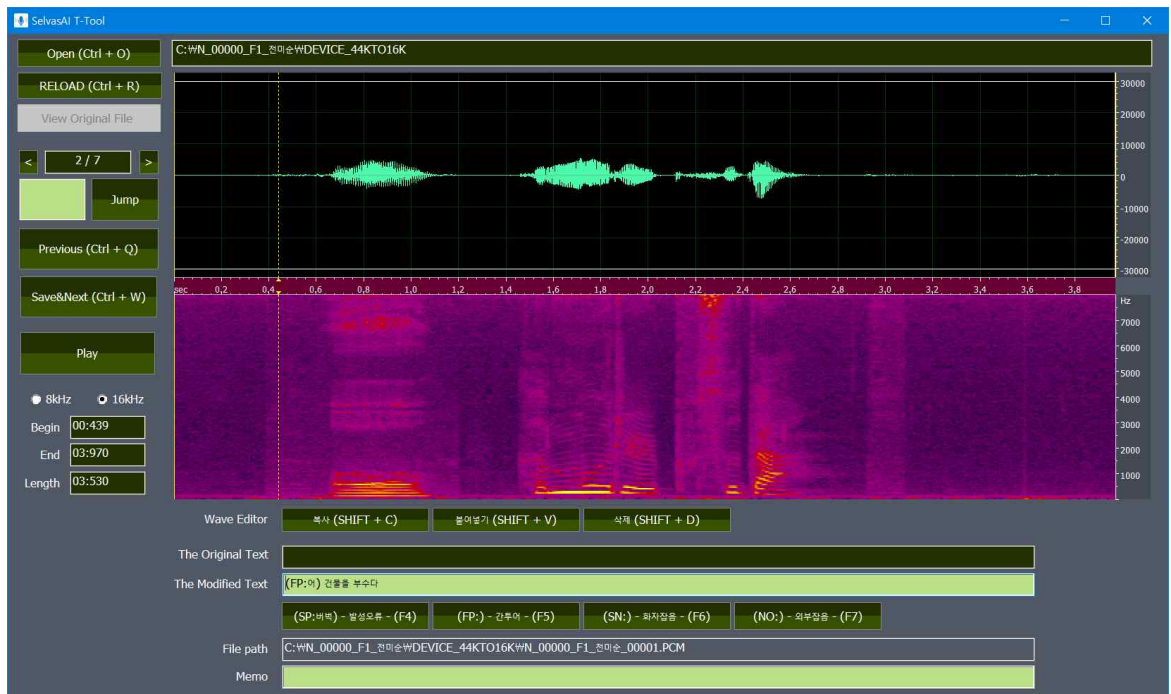
- 1) 입술 소리, 숨소리, 웃음 소리 는 Speaker Noise (SN:) 추가
  - 입술 소리나 숨소리는 대부분 녹음실 환경에서 근접 마이크로 녹음시에만 녹음됨  
예) (SN:) 건물을 부수다 (SN:)



<잡음 및 Filter정보 추가 - Speaker Noise (SN:)>

2) 문장 시작 전에 습관적으로 발성하는 뭐, 음, 어 같은 발화가 녹음된 경우 Filler Noise (FP:음), (FP:어) 추가

예) (FP:어) 건물을 부수다

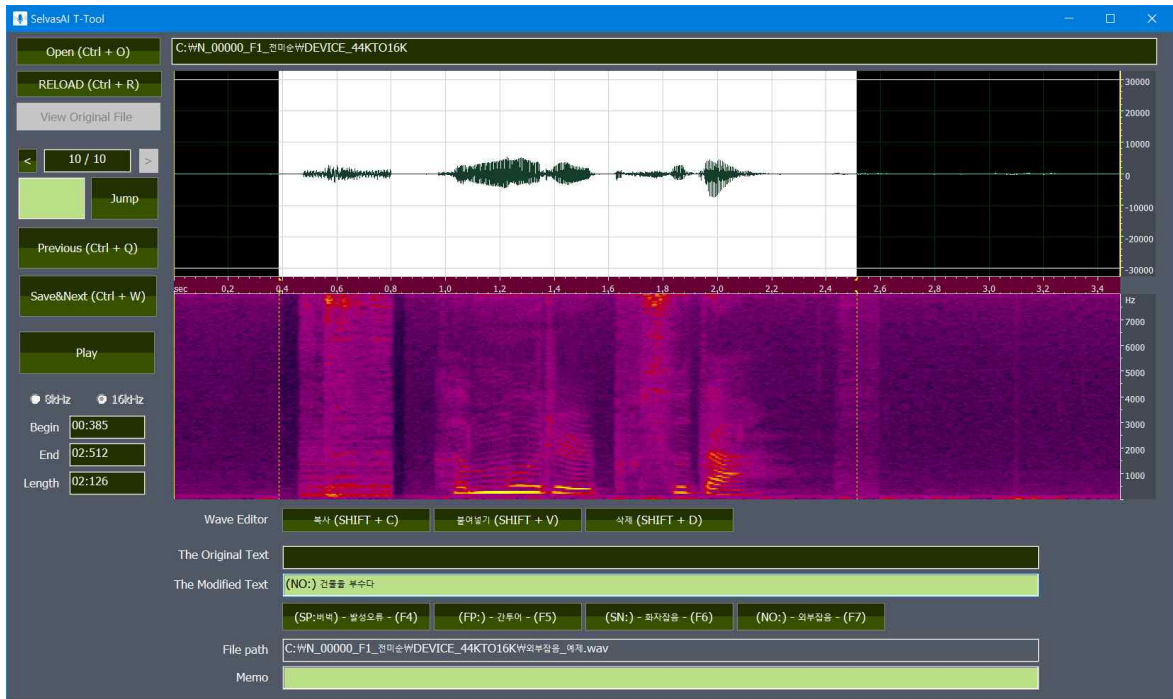


<잡음 및 Filter정보 추가 - Filler Noise (FP:어)>

3) 발화자 이외의 녹음 중외부 잡음이 들어 온 경우, (NO:) 추가

예) (NO:) 건물을 부수다





#### <잡음 및 Filter정보 추가 - 외부 잡음(NO:)>

- 파일별 전사 결과 저장
  - 음원 저장 폴더와 동일한 이름\_OK 폴더 생성

DEVICE_44KTO16K	2020-09-28 오전 8:50	파일 폴더
DEVICE_44KTO16K_OK	2020-09-28 오전 8:51	파일 폴더

#### [그림 11] 저장 폴더(\_OK) 생성

- \*\*\_OK 폴더 아래에 음원과 전사 정보 파일(.TRS) 저장

[Text Information]  
 The Original EPD Start=0  
 The Original EPD End=348  
 The Original Text=건물을 부수다  
 The Original Sample Rate=16000  
 The Modified EPD Start=0  
 The Modified EPD End=348  
 The Modified Text=(NO:) 건물을 부수다  
 Memo=  
 Voice Type=

#### <전사 정보 파일(.TRS) 예시>

#### ○ 어노테이션/라벨링 도구2 (NHN다이렉스트)

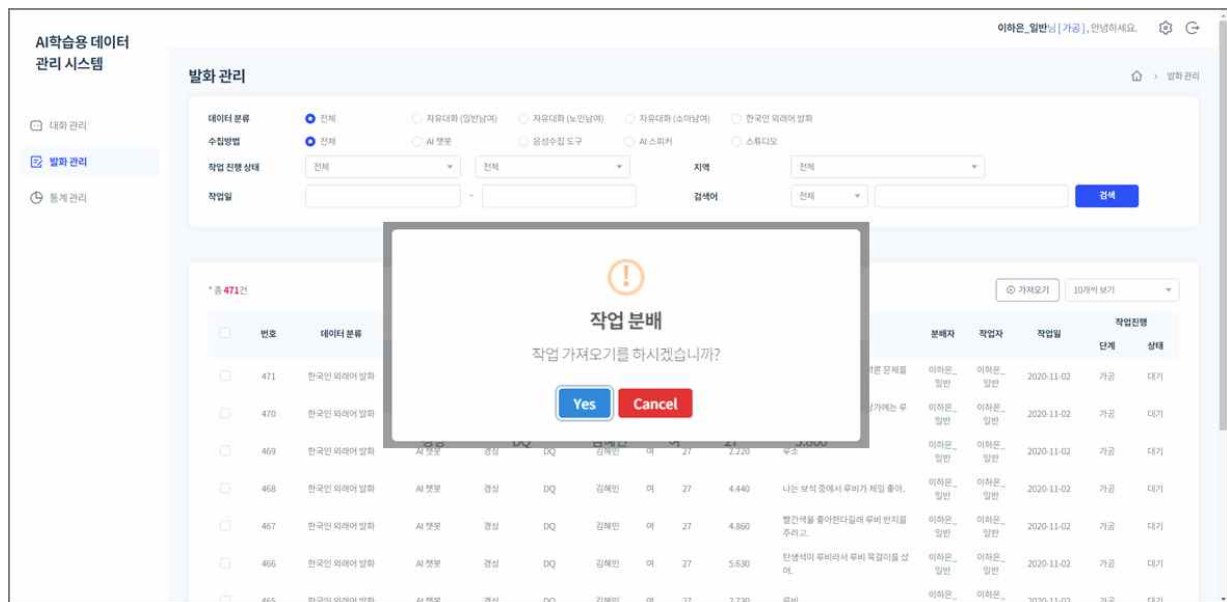
- 개요
  - 데이터 가져오기 및 작업 분배 기능을 통한 검수 데이터 배분
  - 다수의 검수자가 동시에 병렬 검수 작업 가능

- 검색 및 필터링 기능을 통한 검수 대상 발화 식별
- 음성데이터 재생 및 가공(전사) 문서 비교하여 수정
- 반려 기준에 따라 검토하고, 기준을 충족하지 못할 경우 반려 처리

• 전사도구(AI학습용 데이터 관리 시스템) 사용법

- 발화 가져오기

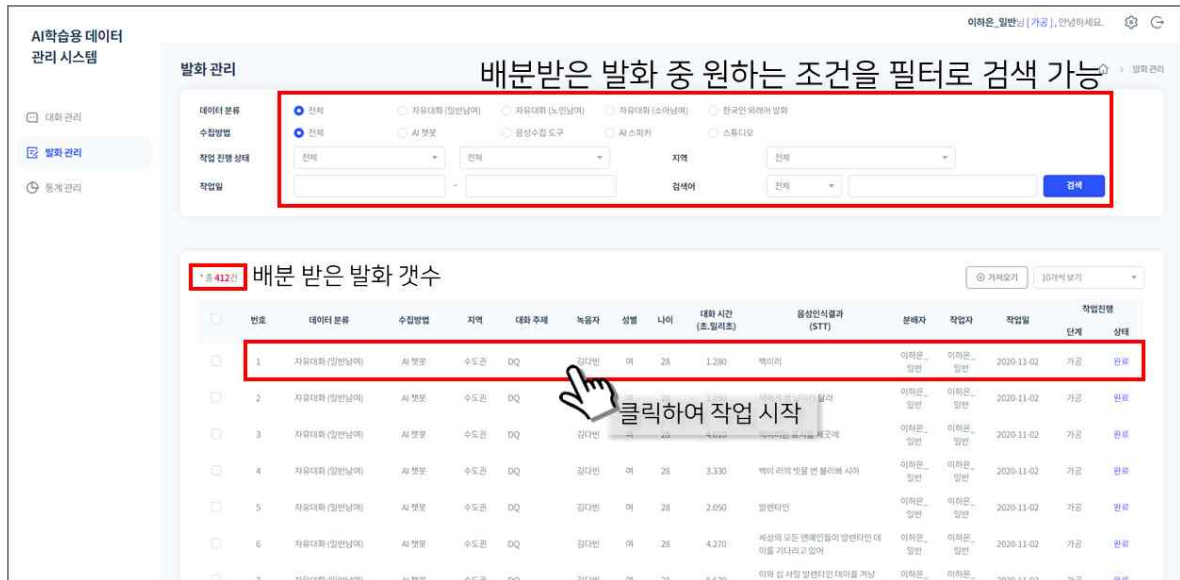
- 가져오기시 10분(약 150 문장)분량 데이터 배분
- 최대 10분 분량만 배분 가능하며 작업 완료 전 추가로 가져오기 누를 시 작업 진행된 분량 만큼만 추가



<발화관리 - 발화 가져오기 및 작업분배>

- 작업 시작

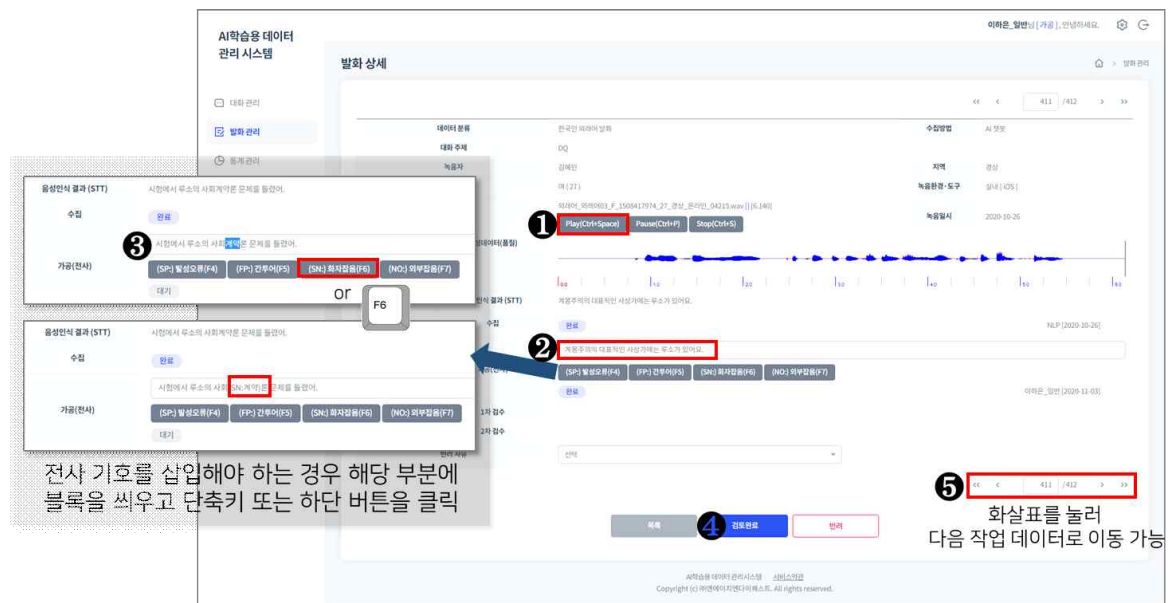
- 배분 받은 발화 중 원하는 조건을 필터로 검색 가능
- 배분 받은 발화 개수 확인 및 발화 선택하여 가공 작업
- 가공 대기 상태로 검색 시 작업 대기 중인 데이터만 검색
- 작업진행 단계는 '분배 후 작업전: 가공 대기', '검토 완료: 가공 완료'로 구분



<발화관리 - 발화 배분>

#### - 가공(전사)

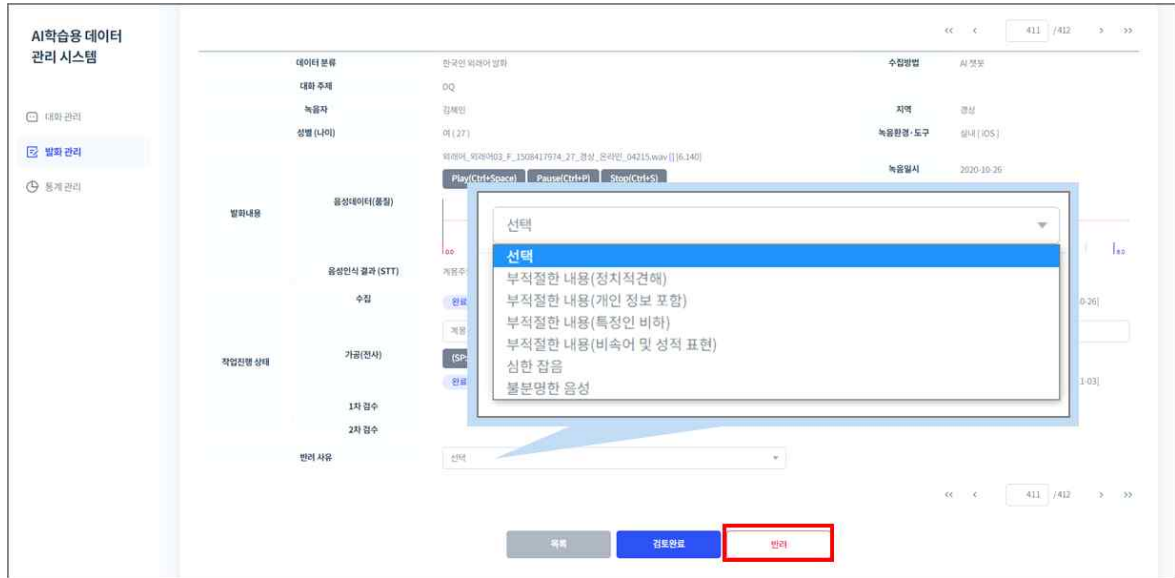
- Play 버튼(Ctrl+Space)을 눌러 녹음된 음성데이터를 듣고 가공(전사) 문장과 비교하여 수정
- 전사 기호를 삽입해야 하는 경우 해당 부분에 블록을 씌우고 단축키 또는 하단 버튼을 클릭하여 처리
- 검토완료 버튼 클릭하여 저장
- 화살표를 눌러 다음 데이터로 이동



<발화관리 - 가공(전사)>

#### - 반려

- 반려 항목의 경우 반려 사유를 선택하고 검토 완료 버튼이 아닌 반려 버튼 클릭
- 반려 기준은 '정치적 견해 포함', '개인 정보 포함', '특정인 비방', '비속어 및 성적 표현', '심한 잡음', '불분명한 발음' 등이 있음

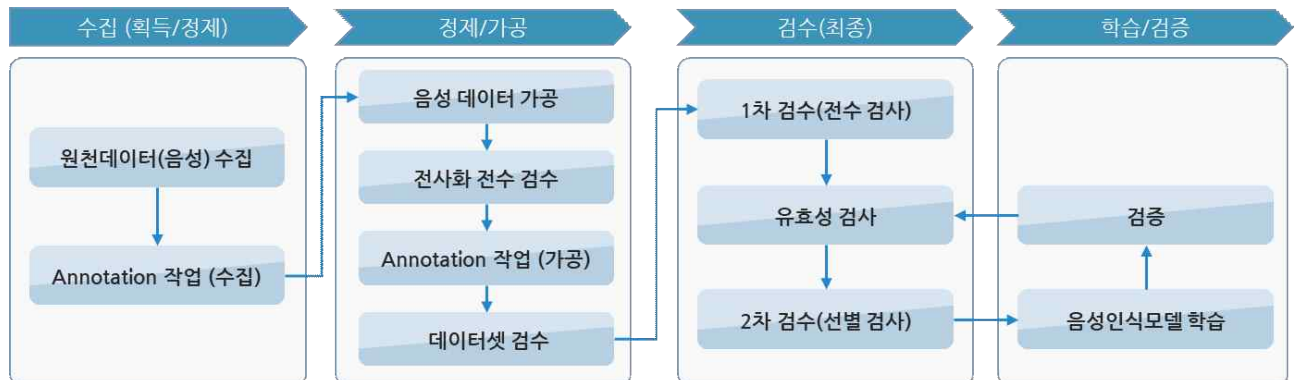


<발화관리 - 반려>

## 2.5 검수

### 2.5.1 검수 절차

#### ○ 검수 프로세스



세부	수집	정제/가공	검수(최종)	학습/검증
검수 단계	<ul style="list-style-type: none"> <li>PCM 형태의 음성파일 수집</li> <li>Annotation 작업(수집)을 통해 노이즈 정도 체크</li> <li>손상된 음성, 인식하기 어려운 음성 데이터 오류 체크 및 제외</li> <li>분류기준(카테고리)별 음성 분류 유무 체크</li> <li>수집 데이터 작업자 자체 및 Cross Check</li> </ul>	<ul style="list-style-type: none"> <li>음성 데이터 텍스트화 작업</li> <li>전사화 전수 검수를 통해 음성 파일 및 전사 텍스트 정확성 체크 및 민감 이슈(정치, 개인정보 등) 발언 저장 도구 Tool 활용하여 제외 처리</li> <li>전사 데이터 라벨링 작업을 통해 분류 기준 별 데이터 매칭 여부 체크</li> <li>전사 데이터에 대한 자체 및 크로스 검증을 통해 어노테이션 검증 결과 일치율 96% 이상이면 과반수에 의해 True 처리</li> </ul>	<ul style="list-style-type: none"> <li>가공 완료된 최종 산출물 (학습데이터) 전수 검수</li> <li>라벨링 누락 및 정확성 체크</li> <li>전체 데이터셋 기준 (음성 텍스트, 메타정보) 통계 기반 체크</li> </ul>	<ul style="list-style-type: none"> <li>분석 모델 Validation Test를 통한 성능 평가 및 학습데이터 테스트</li> </ul>

#### ○ 검수 단계에 따른 세부 내용

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) 데이터 라벨링 작업 데이터 전사화 작업 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>

#### ○ 크로스 체크를 통한 검수 일관성 유지

- 검수 단계에서 검수 작업자간의 일관성을 유지하기 위해 크로스 체크를 통한 검수 진행
- 검수 작업 시작 시 별도의 교육을 통해 검수기준을 고지하고, 예시를 통한 검수기준 체득
- 검수 작업에 애매한 부분을 별도로 체크하여 정리하고, 주기적인 작업내용 및 크로스체크 사항에 대한 회의를 통해 검수 기준을 확립하고 검수 일관성 유지
- 음성 데이터의 심한 잡음에 대한 오류 체크 작업 시, 각 검수 작업자 별 검수기준이 상이할 수 있으므로, 잡음과 그 그정도에 대한 정의를 실제 음성 데이터 사례를 중심으로 교육 및 회의를 진행
- 검수 작업 중 잡음 등에 대해 검수기준이 모호한 부분 등은 해당 음성 데이터 및 내용을 공유하여 각 작업자 간의 검수 일관성 유지
- 음성 데이터에 포함된 비속어 및 은어 중 실제 자유 대화에서 사용될 수 있다고 판단되는 경우는 제외처리 하지 않고 유지
- 비속어 및 음성 검수 기준을 정의하여 검수 작업자 교육을 진행하고, 작업자 별로 비속어 및 은어 통용범위에 대해 검수 기준이 다를 수 있으므로, 회의 및 예시를 통한 검수 일관성 유지

#### 2.5.2 검수 기준

단계	검수 역할	검수 내용									
수집	원천 데이터(음성) 수집	- 분류기준(카테고리)별 음성 분류 유무 체크 - 클라핑, frame drop 등 손상된 음성신호 제외 처리 체크									
	어노테이션(수집)	- 심한 잡음으로 발생한 오류 체크 - 수집 단계 작업자의 자체 및 크로스 검증									
정제 /가공	음성 데이터 가공	- 음성 데이터 텍스트화 작업									
	전사화 전수 검수	- 문장별 전사 데이터 일치성 체크 - 음성 파일 및 전사 텍스트 정합성 체크 - 민감 이슈(정치, 개인정보 등) 발언 저작 도구 Tool 활용하여 제외 처리									
	어노테이션(가공)	- 전사 데이터 라벨링 작업 - 분류 기준 별 전사 데이터 매칭 여부 체크 예) 분류 기준별 표준화에 대한 규칙 체크 및 수정 <table border="1"><tr><td>분류 기준</td><td>수집 데이터</td><td>표준화 데이터</td></tr><tr><td>성별 표준화</td><td>'남', '남자', 'M' ...</td><td>'남자'</td></tr><tr><td>연령 표준화</td><td>'10','10대','11' ...</td><td>'10대'</td></tr></table> - 분류 기준에 따른 어노테이션 정보(메타 정보) 매핑, 매핑 값 검수 예) 어노테이션 검증 결과 일치율 96% 이상이면 과반수에 의해 True 처리	분류 기준	수집 데이터	표준화 데이터	성별 표준화	'남', '남자', 'M' ...	'남자'	연령 표준화	'10','10대','11' ...	'10대'
	분류 기준	수집 데이터	표준화 데이터								
성별 표준화	'남', '남자', 'M' ...	'남자'									
연령 표준화	'10','10대','11' ...	'10대'									
데이터셋 검수	- 라벨링 누락 및 정합성 체크 - 문장별 전사 데이터 일치성 체크 - 전사된(음성 -> 텍스트) 데이터에 대한 자체 및 크로스 검증										
검수 (최종)	데이터셋 전수 검수	- 구축된 데이터셋 기반 전수 검사 예) 관리자의 통계기반 데이터 셋 체크 - 구축 데이터셋 라벨링 누락 및 정합성 체크 - 구축 데이터셋 어노테이션 누락여부 확인 - 학습데이터 정확도 확인									
활용	Validation Test	- 학습 데이터 기반 서비스 확인									
	학습데이터 테스트	- 모형 성능 평가 지표 기반 분석 모델 성능 평가									

### 2.5.3 검수 조직(필요 시 작성)

//가공한 데이터를 검수하는 조직의 구성과 각 구성원의 역할별 책임과 권한을 구분하여 작성합니다.

//상호 검수 또는 다수결 합의 검수를 진행하는 경우 그 절차를 위한 조직 구조가 드러나도록 작성합니다.

//데이터를 절차에 맞게 검수하기 위해 실시하는 교육 및 훈련 계획을 작성합니다.

### 2.5.4 검수 도구

#### ○ 검수도구 1: AI 학습용 데이터 관리 시스템 (통합 도구)

##### • 대화 관리

- AI 음성 녹음 도구에서 자동으로 연계되어 업로드된 파일의 그룹과 관리도구의 '업로드'버튼을 통해 별도로 파일 업로드한 파일의 그룹 목록 제공
- 데이터 분류, 대화 주제, 총 참여자, 진행상태, 등록일 등의 정보 제공 및 필터링을 통한 대화 검색

## 기능 제공

- 각 데이터 그룹 별로 발화목록을 확인 및 삭제 기능 제공

## ① [검색조건]

## - 데이터 분류

- 자유대화(일반남여), 자유대화(노인남여), 자유대화(소아남여, 유아 등 혼합), 한국인 외래어 발화

## - 진행상태

- 수집: 작업자가 대화주제에 대해 음성 녹음을 완료한 상태)
- 가공: 가공 담당자가 발화목록 검토 진행 중
- 가공 완료: 가공 담당자가 발화목록 검토 완료
- 1차 검수: 1차 검수자가 발화목록 검토 진행 중
- 1차 검수 완료: 1차 검수자가 발화목록 검토 완료
- 2차 검수: 2차 검수자가 발화목록 검토 진행 중
- 2차 검수 완료: 2차 검수자가 발화목록 검토 완료

## - 검색어: 대화주제, 사용자ID

## ② [발화목록] 버튼 클릭하면 발화목록 관리 화면으로 이동

## ③ 오프라인으로 작업한 데이터를 일괄 업로드할 수 있는 화면으로 이동 (다음 화면)

## • AI 데이터 파일 업로드

- 대화 관리 화면에서 '업로드'버튼 클릭 시 레이어 팝업으로 파일 업로드 화면 실행
- 각 데이터 분류 및 대화 주제 정의 후 대화목록 엑셀파일 및 대화 음성 파일 일괄 업로드

① 대화목록 엑셀 양식을 다운로드해서 발화목록 엑셀 작성 (엑셀 항목)

- 음성파일명
- 전사 텍스트
- 지역
- 성별
- 연령

② 데이터분류 선택

- 대화 주제 입력
- 작성한 엑셀 파일 첨부
- 대화 목록의 실제 음성파일들을 첨부

• 발화 관리

- 대화 관리 화면에서 '발화목록'버튼 클릭 시 각 데이터 그룹 내의 발화 건수 확인 및 검수 가능
- 발화내용 및 STT 일치 여부, 발화자ID, 작업 관리 진행 여부 등 확인 가능하고, 각 발화 별 '검수'버튼을 통해 검수 진행



인공지능 데이터 구축·활용 가이드라인 양식 v0.4

관리자님 [가공], 안녕하세요. < >

인공지능 데이터 관리 시스템

대화 관리 발화 관리 통계 관리 회원 관리 코드 관리

발화 관리

데이터 분류 ☒ 전체 ☐ 자유대화 (일반남여) ☐ 자유대화 (노인남여) ☐ 자유대화 (소년남여) ☐ 한국어 외래어 발화

수집방법 ☒ 전체 ☐ AI 챗봇 ☐ 음성수집 도구 ☐ AI 스피커 ☐ 소용도

작업 진행 상태   지역

작업일  ~  검색어  검색

\* 총 266건

중 작업 분배 발화 등록 음 선택 삭제 가져오기 다운로드 10개씩 보기

번호	데이터 분류	수집방법	지역	대화 주제	녹음자	성별	나이	대화 시간 (초:밀리초)	음성인식결과 (STT)	분배자	작업자	작업일	작업 진행 단계	상태	
<input type="checkbox"/>	266	자유대화 (일반남여)	AI 챗봇	수도권	DQ	최현욱	여	30	1.760	안녕하세요	최현욱	2020-10-21	수집	완료	
<input type="checkbox"/>	265	자유대화 (일반남여)	AI 챗봇	수도권	DQ	최현욱	여	30	2.140	날씨 테스트 중	이하은	최현욱	2020-10-21	수집	완료
<input type="checkbox"/>	264	자유대화 (일반남여)	AI 챗봇	수도권	DQ	최현욱	여	30	1.950	관료 테스트	이하은	최현욱	2020-10-21	수집	완료
<input type="checkbox"/>	263	자유대화 (일반남여)	AI 챗봇	수도권	DQ	최현욱	여	30	2.690	한글 테스트	이하은	최현욱	2020-10-21	수집	완료
<input type="checkbox"/>	262	자유대화 (일반남여)	AI 챗봇	수도권	DQ	최현욱	여	30	2.410	플루티스트	이하은	최현욱	2020-10-21	수집	완료
<input type="checkbox"/>	261	자유대화 (일반남여)	AI 챗봇	수도권	DQ	최현욱	여	30	2.410	별도 테스트 중	이하은	최현욱	2020-10-21	수집	완료
<input type="checkbox"/>	260	자유대화 (일반남여)	AI 챗봇	수도권	DQ	최현욱	여	30	1.760	반려 테스트	이하은	최현욱	2020-10-21	수집	완료
<input type="checkbox"/>	259	자유대화 (일반남여)	AI 챗봇	수도권	DQ	최현욱	여	30	1.950	안녕하세요	관리자	관리자	2020-10-21	가공	대기
<input type="checkbox"/>	258	자유대화 (일반남여)	AI 챗봇	수도권	DQ	최현욱	여	30	1.670	안녕하세요	관리자	관리자	2020-10-21	가공	대기
<input type="checkbox"/>	257	자유대화 (일반남여)	AI 챗봇	수도권	DQ	최현욱	여	30	2.410	안녕하세요	관리자	관리자	2020-10-21	가공	대기

1 2 3 4 5 6 7 8 9 10 > >>

## ① 진행상태

- 수집완료: 데이터 구축자가 대화주제에 대해 음성 녹음을 완료한 상태
- 가공완료: 가공자가 각 발화에 대해 가공을 완료한 상태

## ② 작업관리

- [검수]: 발화목록이 작업 완료인 상태로 [검수] 버튼 클릭하면 검토화면으로 이동
- 가공완료: 가공자가 검토 완료한 상태

## • 발화 검토

- 발화관리 메뉴에서 '검수'버튼 클릭 시 발화의 기본 정보(발화자ID, 지역, 성별, 연령) 확인 및 발화의 음성 데이터를 바로 재생하여 전사텍스트와 비교하여 검토 가능
- 발화 음성과 전사데이터가 동일 한 경우, 일치 선택하여 가공완료 처리하고, 불일치하는 경우 불일치 선택하여 가공완료 처리
- 일치 선택한 경우 1차 검수자 및 2차 검수자를 통해 검수 진행되고, 불일치 발화는 검수 단계로 진행되지 않음
- 검토 권한(가공, 1차 검수, 2차 검수)에 따라 발화 검토 진행

250 / 257

데이터 분류	한국인 외래어 발화		수집방법	음성수집 도구
대화 주제	생활		지역	수도권
녹음자	김하나	녹음환경·도구	실내 [ ANDROID ]	
성별 (나이)	남 (33)	녹음일시	2020-10-20	

2

발화내용

음성데이터(품질)

소아남여\_소아남1401\_M\_1455308554\_2\_5\_수도권\_온라인\_50002.wav [ 2.470 ]

Play
Pause
Stop

음성인식 결과 (STT)

바닷가에 환어요

가공(전사)

동화도우미

(SP-) 발성오류(F4)
(FP-) 간투어(F5)
(SN-) 화자잡음(F6)
(NO-) 외부잡음(F7)

완료

작업진행 상태

1차 검수

2차 검수

반려 사유

선택

반려

250 / 257

목록
검토완료

### ① 발화 순서

- [목록] 버튼을 클릭하여 발화목록 화면으로 다시 이동하지 않고, 발화검토 화면의 좌, 우 버튼 클릭으로 계속 발화 검토 가능

### ② 발화내용 검토

- 검토자(가공자)는 발화내용의 음성과 전사 텍스트를 각각 비교 검토하여 정확하게 일치하면 [가공 완료] 버튼을 클릭하여 검토 완료

### ③ 작업상태에 따라 다른 버튼명을 제공하고, 버튼 클릭 시 가공 및 검토 완료

- 작업완료: [ 가공 완료 ]
- 가공완료, 1차 검수 완료: [ 검토 완료 ]

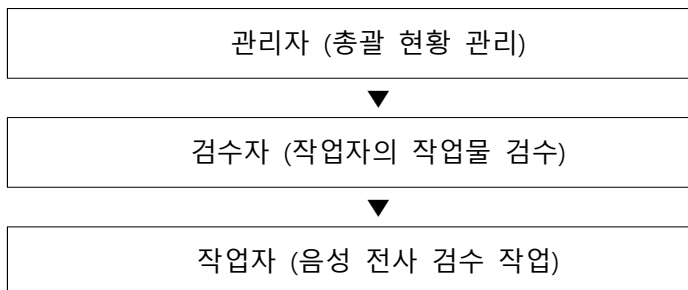
## ○ 검수도구 2: 클라우드소싱 및 웹 기반 작업 도구

- 기본적으로 Public 클라우드소싱 / Private 클라우드소싱 작업자 모두 범용적으로 사용이 가능하며 추후 플랫폼으로써 확장성을 고려한 웹 기반 작업 도구 구축 완료

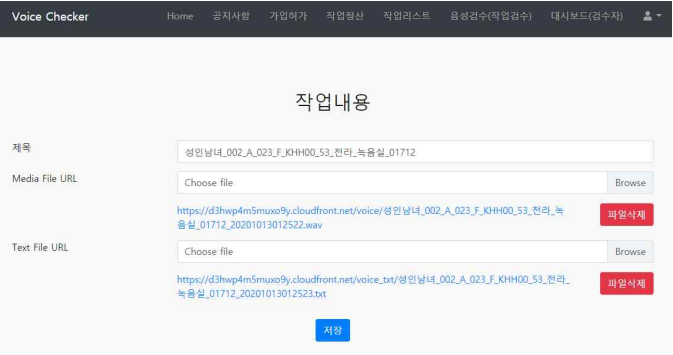
Voice Checker								
Home   공지사항   가입허가   작업정산   작업리스트   음성검수(작업검수)   대시보드(검수자)								
Show 10 entries		Search: <input type="text"/>						
#	제목	작성자	생성일	수정일	상세			
1	DB 이용관련 문의	admin	2020-10-07	2020-10-07	<a href="#">상세</a>			
Showing 1 to 1 of 1 entries					Previous 1 Next			

&lt;웹 기반 작업 도구 메인화면&gt;

- 접근성 및 편리성이 강화된 웹 기반 플랫폼을 구축하여 언제 어디서든 장소에 구애받지 않고 데이터를 검수할 수 있는 작업환경을 구축하고 작업자, 작업자를 검수하는 검수자, 총 현황을 관리하는 관리자의 3단계 유저 레벨 작업 체계를 구축하여 사업 기간 내에 과업을 완료할 수 있도록 함

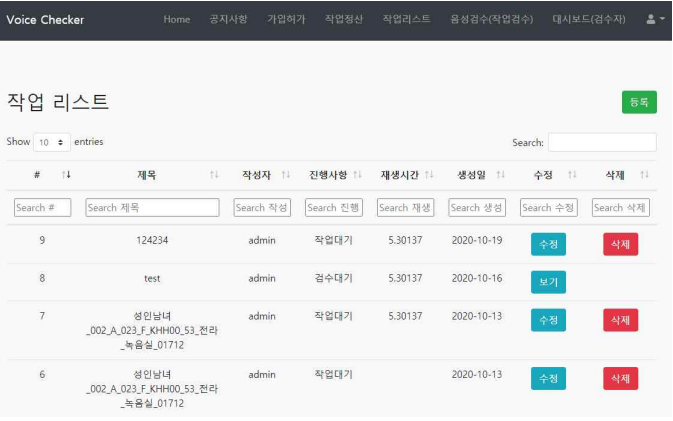


- PCM, WAV, MP3 등 다양한 확장자를 가진 음성파일들을 품질 저하 없이 HTML5 웹 기반에서 구동되도록 하는 모듈 탑재



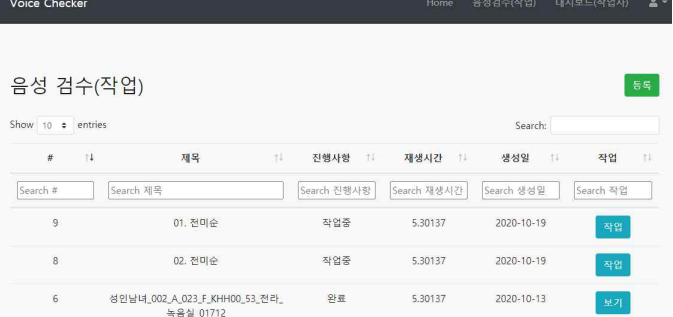
수집 된 PCM,  
전사 된 TXT  
파일 등록





업로드된 파일은  
스크립트를 통해  
웹에서 재생 가능한 형태로  
자동 변환된 후  
작업리스트에 등록





작업자들은 등록된 작업물을  
보며 검수 작업 실시

- 작업자/검수자/관리자의 검수 작업 프로세스
  - 작업자는 1)등록된 작업을 선택하고 2)작업 화면에서 작업한 이후 3)저장 및 제출을 함으로써 1개 작업 프로세스 종료 이후 해당 프로세스 반복
  - 검수자는 작업자별 제출한 작업물을 확인한 후 완료/반려 여부 결정
  - 관리자는 검수자가 검수를 완료한 작업물에 대해 작업자에게 정산 실시

Voice Checker

Home 음성검수(작업) 대시보드(작업자)

음성 작업

Title

01. 전미순

Media

▶ 0:00 / 0:05

0.5 1 1.5 2 3

음성 Text

건물을 부수다

검수 Text

건물을 깨부다가 말았다

작업 비교

건물을 부수깨부다가 말았다

저장

제출

#### \*작업자의 작업 화면

- 등록된 음성을 재생하며, 전사된 파일을 보고 제대로 됐는지 검수 실시
- 검수 중 잘못 된 전사작업 확인시 수정 작업 실시
- '작업 비교'란에서 수정된 부분 자동표시

Voice Checker

Home 공지사항 가입하기 작업정산 작업리스트 음성검수(작업검수) 대시보드(검수자)

작업 내용

제목

test

Media URL

https://d3hwp4m5muxo9y.cloudfront.net/voice/test\_20201016120748.wav

▶ 0:00 / 0:05

Text URL

https://d3hwp4m5muxo9y.cloudfront.net/voice\_txt/test\_20201016120749.txt

얼마 전에 그 집 아들 생일이었지?

작업 내용

얼마 전에 그 집 아들 생일이었지?

작업 비교

얼마 전에 그 집 아들 생일이었지?

승인

반려

검수자의  
승인 / 반려  
작업 창

Voice Checker

[Home](#)
[공지사항](#)
[가입하기](#)
[작업정산](#)
[작업리스트](#)
[음성검수\(작업검수\)](#)
[대시보드\(검수자\)](#)

작업 정산

검수 중 개수 : 5 | 검수 중 시간 : 26.40686

Show 10 entries

#	제목	작업자	진행사항	재생시간	생성일	정산	정산
9	01. 전미순	김진희	작업중	5.30137	2020-10-19	미정산	
8	02. 전미순	김진희	작업중	5.30137	2020-10-19	미정산	
7	test	cc	검수대기	5.30137	2020-10-16	미정산	
6	성인남녀_002_A_023_F_KHH00_53_전라_녹음실_01712	김진희	완료	5.30137	2020-10-13	정산	<input type="button" value="정산"/>
5	02. 전미순		완료	5.30137	2020-10-07	정산	<input type="button" value="정산"/>
4	02. 전미순		완료	5.30137	2020-10-05	미정산	<input type="button" value="정산"/>

관리자의  
정산 / 미정산  
처리 작업 창

- 클라우드소싱 및 웹 기반 작업 도구의 장점

- 다수의 작업자가 퍼블릭 클라우드소싱 형태로 참여가 가능하고, 가입신청 이후 승인된 작업자에 한해 작업 권한을 제공하므로, 제한된 퍼블릭 클라우드소싱 형태로 작업이 가능한 범용적인 플랫폼 형태의 도구
- PCM, WAV, MP3 등 다양한 형태의 음성파일을 웹에서 재생될 수 있도록 자동화시키는 모듈을 탑재하여, 음성검수 작업자가 작업 진행 시 자신이 작업 내용을 '작업비교'를 통해 손쉽게 확인할 수 있는 스크립트가 짜여 있어, 활용성 및 범용성이 우수

### 2.5.5 기타 품질관리 활동(필요 시 작성)

-

## 2.6 활용

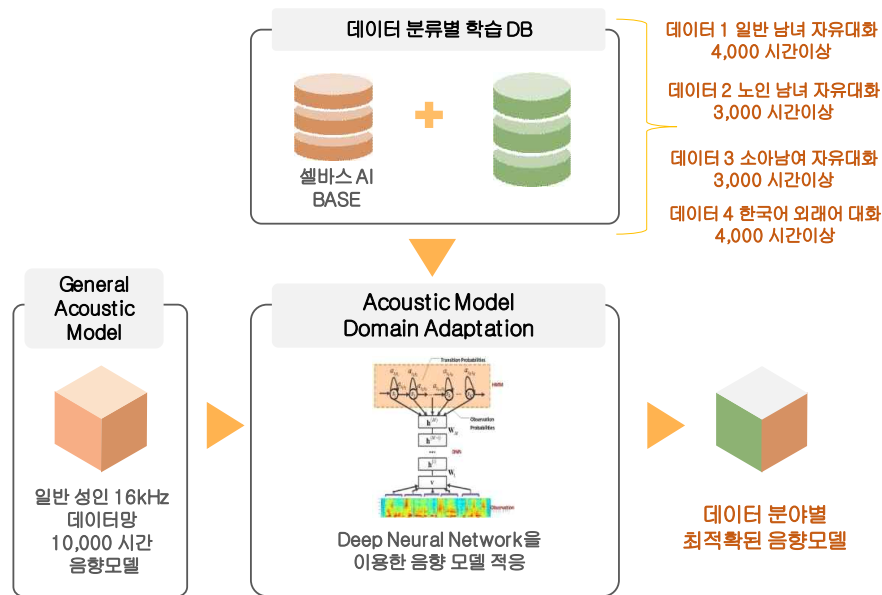
### 2.6.1 활용 모델

#### 2.6.1.1 모델 학습

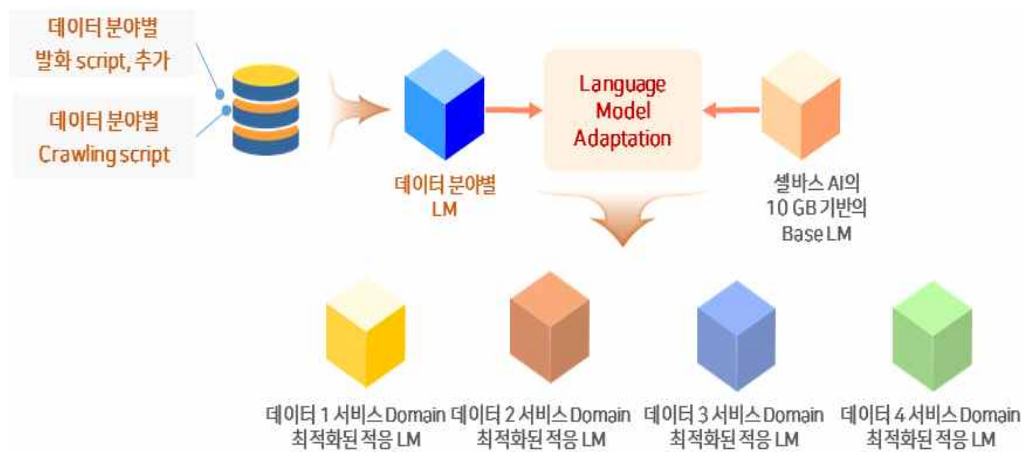
- 모델학습의 의의
  - 수집된 분야별 데이터(일반남녀, 노인남녀, 소아남녀, 한국인 외래어 발화)에 대해 데이터 품질검증을 하는 방식으로 셀바스 AI BASE 모델 분야별 적응 학습을 수행하고, 인식 성능 향상됨을 확인하여 데이터의 유효성을 검증
- 데이터 1 : 자유대화(일반남녀)
  - 응용 AI 스피커로 녹음된 1,000개의 TEST DB 생성
  - 셀바스 AI BASE 인식엔진으로 음절단위 인식 성능 평가
  - 데이터 1에서 수집된 다양한 환경의 학습 DB ( Android, iOS, PC, AI 스피커 )를 이용하여 음향 모델 학습, 서비스 도메인의 문장 기반의 언어모델 학습
  - AI 스피커 음원이 포함된 적응 엔진 으로 음절 단위 인식 성능 평가
  - BASE / 적응 모델 간의 음성인식 성능 확인
  - 적응 모델에 AI 스피커 음향모델이 적용함에 따라 인식 성능 향상 확인을 통해 수집된 데이터의 유효성 검증
- 데이터 2 : 자유대화(노인남녀)
  - Android, iOS 단말로 녹음된 1,000개의 TEST DB 생성
  - 셀바스 AI BASE 인식엔진으로 음절단위 인식 성능 평가
  - 데이터 2 에서 수집된 다양한 환경의 학습 DB (Android, iOS, PC, AI 스피커)를 이용하여 음향 모델 학습, 서비스 도메인의 문장 기반의 언어모델 학습
  - 노인 음원 음원이 포함된 적응 엔진 으로 음절 단위 인식 성능 평가
  - BASE / 적응 모델 간의 음성인식 성능 확인
  - 적응 모델에 노인 음원 음향모델이 적용함에 따라 인식 성능 향상 확인을 통해 수집된 데이터의 유효성 검증
- 데이터 3 : 자유대화(소아남녀)
  - Android, iOS 단말로 녹음된 1,000개의 TEST DB 생성
  - 셀바스 AI BASE 인식엔진으로 음절단위 인식 성능 평가
  - 데이터 3에서 수집된 다양한 환경의 학습 DB (Android, iOS, PC, AI 스피커)를 이용하여 음향 모델 학습, 서비스 도메인의 문장 기반의 언어모델 학습
  - 소아 음원 음원이 포함된 적응 엔진 으로 음절 단위 인식 성능 평가
  - BASE / 적응 모델 간의 음성인식 성능 확인
  - 적응 모델에 소아 음원 음향모델이 적용함에 따라 인식 성능 향상 확인을 통해 수집된 데이터의 유효성 검증
- 데이터 4 : 한국인 외래어 발화
  - 외래어 가 포함된 문장으로 구성된 1,000개의 TEST DB 생성
  - 셀바스 AI BASE 인식엔진으로 음절단위 인식 성능 평가
  - 데이터 4에서 수집된 다양한 환경의 학습 DB (Android, iOS, PC, AI 스피커)를 이용하여 음향

모델 학습, 서비스 도메인의 문장 기반의 언어모델 학습

- 외래어 script와 음원이 포함된 적응 엔진 으로 음절 단위 인식 성능 평가
- BASE / 적응 모델 간의 음성인식 성능 확인
- 한국어 발성에 자주 나오지 않는 음소열이 포함된 외래어 음향, 언어모델이 적용함에 따라 인식 성능 향상 확인을 통해 수집된 데이터의 유효성 검증



데이터 분야별 최적화된 음향모델 적응학습



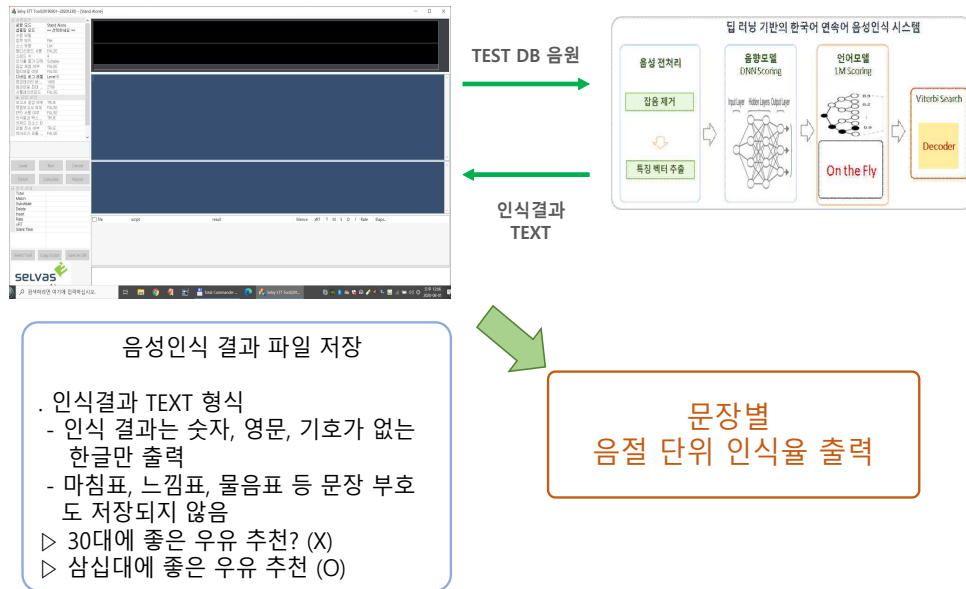
데이터 분야별 최적화된 언어모델 적응학습

#### ○ 한국어 음성인식 성능 평가

##### • 평가 방법

- Base 엔진, 50% AI 학습데이터 DB를 이용한 적응모델 엔진 1, 100% AI 학습데이터 적응모델 엔진 3개의 엔진별로 선별된 TEST DB에 대해 음절 단위로 비교하여 평균 인식률을 산출

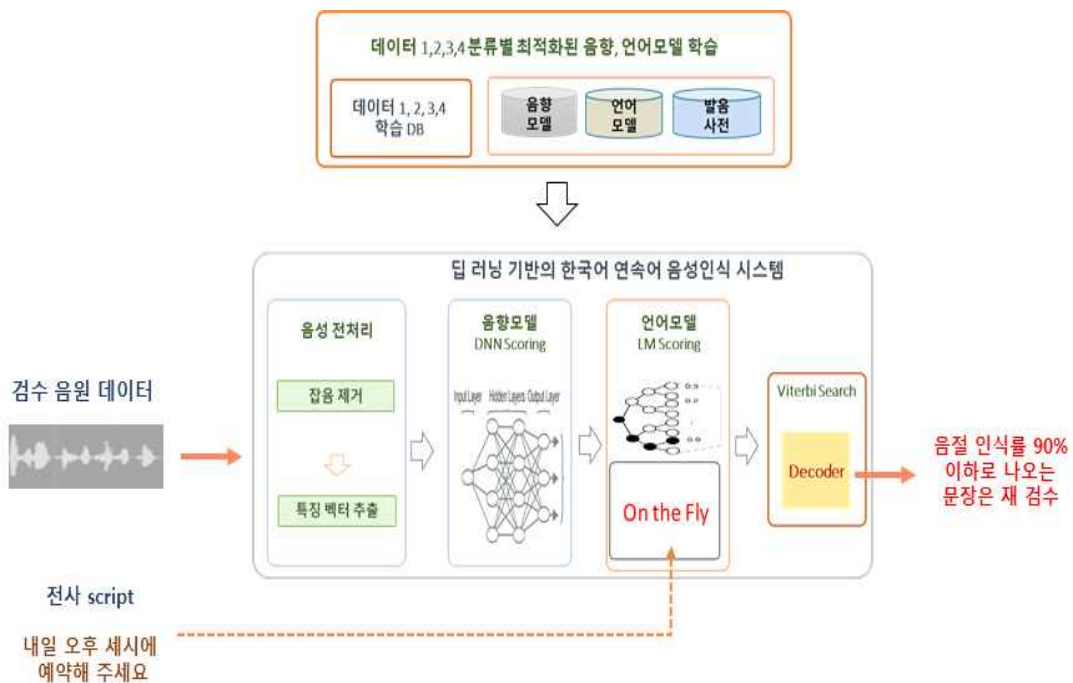




### 데이터 분야별 음성인식 성능 평가

#### ○ 음성인식엔진을 이용한 최종 검수

- 전사, 검수 단계를 수행한 데이터 1,2,3,4 음성학습 DB에 대해 각 데이터별로 최적화된 인식엔진 기반으로 언어모델 On the fly 기법을 도입하여 화자별 전사 script 오류를 2차 검수 수행
- On the fly 기능은 기본 언어모델 인식 네트워크에 전사 script를 인식 후보로 올려놓고 인식을 수행함으로써 즉, On the fly에 입력된 script에 weight를 높여 인식을 하기 때문에 입력 발성과 동일한 전사 script 결과가 나오면 정상이며, 인식 결과가 다르게 나오면 전사 script에 오류가 있다고 판단
- 전사 script가 오류가 있을 경우, 음절 인식률이 90% 이하로 나오는 발화 리스트를 추출하여 재검수



### 최적화된 음성인식엔진을 이용한 최종 검수

- 최종 검수된 AI 데이터 기준으로 최종 검수를 수행하여 음절 인식을 정확도 측정 93% 확보

○ 적응학습된 음성인식엔진의 AI 응용 서비스 적용

■ 본 과제의 분야 1~4에서 수집된 DB를 이용하여 분야별 생성된 음향, 언어모델을 이용하여 AI 응용 서비스 운영할 수 있도록 음성인식엔진을 제공 및 성과활용기간(3년)동안 온라인 시범 서비스 지원

### 2.6.1.2 서비스 활용 시나리오

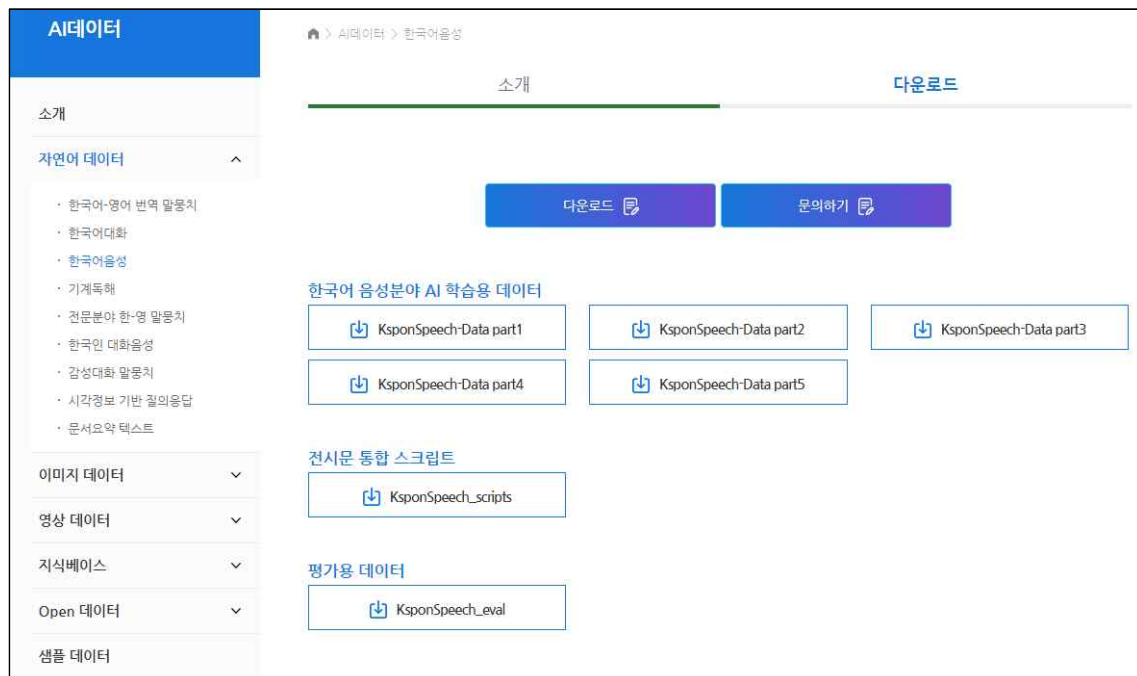
과제명	AI 모델	모델 성능 지표	응용서비스
자유대화 (일반남여)	일반 남여 자유대화 모델	음절 인식을 90% 이상	챗봇 시범 서비스
자유대화 (노인남여)	노인 남여 자유대화 모델	음절 인식을 82% 이상	노인용 챗봇 서비스 "말벗"
자유대화 (소아남여, 유아 등 혼합)	소아/유아 남녀 자유대화 모델	음절 인식을 82% 이상	한국어 교육 서비스 "우리 아이 첫 한글"
한국인 외래어 발화	한국인 외래어 발화 모델	음절 인식을 82% 이상	인공지능 음악추천 서비스 "나만의 뮤직"

- 데이터 1 : 자유대화(일반남녀)
  - 자유대화(일반남여) 데이터를 기반으로 만든 학습 모델에 학습되어 있는 데이터에 한해 사용자가 질문을 하면 매핑 되어 있는 답변을 제공
  - 질문 및 답변 API를 제공하며, Web에서 챗봇을 테스트 해볼 수 있는 클라이언트 환경 제공
  - 사용자가 입력한 메시지에 대한 의도를 파악 후 의도에 맞는 답변을 관리하는 채팅서버 개발
  - 학습용 데이터를 사용한 학습모델의 예측률 향상을 위해 형태소 분석 + CNN을 통한 챗봇 모듈 개발
  - Web기반으로 PC, Mobile에 관계없이 챗봇(프로토타입) 시범 서비스 사용 가능
- 데이터 2 : 자유대화(노인남녀)
  - 노인용 AI로봇의 스피커를 통해 사용자 발화 시, 해당 음성을 TTS로 변환하고 이를 NLP 시스템의 다양한 모듈을 통해 언어이해 및 답변생성을 하여 AI의 로봇의 마이크로 STT 변환하여 답변 출력
  - 단순 질의응답 서비스가 아닌 노인 대상의 말벗 서비스가 중심이므로, 사용자에게 5,000여개의 자연어 목록을 기반으로 다양한 질의 선 제시
  - 질문-답변의 쌍으로 저장된 데이터와, 여러 turn이 반복된 대화를 이용해, 노인용 일상대화 말벗 개발
  - AI 로봇을 활용한 독거노인 돌봄 사업에 활용
  - 발음이 부정확한 노인 사용자의 음성인식 정확도 개선을 통한 챗봇 내 의도 파악 개선 및 사용자 위급상황 인지
- 데이터 3 : 자유대화(소아남녀)

- '우리 아이 첫 한글 학습앱'을 통해 소아, 영유아를 대상으로 음성인식 한글 학습 수행
  - 학습 주제를 선택하고 음성 말하기 연습을 수행하고 이 과정에서 음성인식 기술 활용하여 음성 인식 데이터를 STT로 전사
  - STT로 전사한 데이터로 발음 평가를 수행하고, 음성인식 데이터 분류를 통해 학습콘텐츠를 생성하고 발음평가 수행
  - 최종 학습현황 및 결과를 사용자에게 레포팅 하고 해당 내용은 학습데이터로 저장하여 활용
  - 자유대화(소아남여) 데이터를 활용하여 학습시킨 음성인식 기술을 접목, 발음평가에 활용
  - 영유아, 아동 대상 한국어 학습 프로그램(프로토 타입) 앱 개발
- 데이터 4 : 한국인 외래어 발화
    - '나만의 뮤직'어플리케이션을 음성명령 및 음성인식(STT)를 통해 음성 문장 내 개체명 인식하고 음악 검색
    - 메타정보를 통해 음악을 검색하여 답변 발화를 생성 및 음성합성(TTS)을 통해 음악 추천
    - 한국인 외래어 발화 데이터를 활용하여 학습시킨 음성인식 기술을 접목, 영어표현의 콘텐츠의 제목, 인명, 지명 등의 인식률을 향상
    - 다양한 형태로 발음되어 이형태로 텍스트 변환되는 외래어/외국어에 대한 개체명(제목, 인명, 지명)인식률 향상

## 2.6.2 데이터 제공

### ○ 데이터 제공 방안



<AI 허브의 자연어 데이터 - 한국어 음성 다운로드 페이지>

- 한국정보화진흥원에서 운영하는 AI 허브(<https://aihub.or.kr/>)를 통해 인공지능 데이터를 제공

- AI데이터 > 자연어데이터 > 한국어음성 메뉴 또는 신규 메뉴 추가하여 해당 데이터 정보 및 다운로드 제공
- AI 허브 내 별도의 신청절차에 따라 사용목적 적합하다고 판단될 경우 해당 데이터 제공

#### ○ 데이터 제공 정보

- 데이터 이름: 자유대화 AI 데이터
    - 자유대화 (일반남여)
    - 자유대화 (노인남여)
    - 자유대화 (소아남여, 유아 등 혼합)
    - 한국인 외래어 발화
  - 활용분야
    - 연구분야 : 음성인식, 음성언어처리, 자연어처리, 한국어 음성언어연구, 신호처리 등
    - 산업분야 : 온/오프라인 기반의 음성인식, AI비서, Voice BOT, Voice Command & Control, AI 로봇, 음성인식기반 키오스크
  - 데이터 요약
    - 자유대화 (일반남여)
      - 10대에서 50대 사이의 일반인 남녀의 발화 데이터
      - 녹음 인원 2,000명 이상, 4,000시간 음성 데이터
    - 자유대화 (노인남여)
      - 60세 이상의 남녀 발화 데이터
      - 녹음 인원 1,000명 이상, 3,000시간 음성 데이터
    - 자유대화 (소아남여, 유아 등 혼합)
      - 3세~6세, 7~10세 연령의 남녀 발화 데이터
      - 녹음 인원 1,000명 이상, 3,000시간 음성 데이터
    - 한국인 외래어 발화
      - 녹음 인원 2,000명 이상 4,000시간 음성 데이터
      - 한국인이 발성한 외래어가 포함된 음성 데이터
  - 데이터 출처
    - 신규 제작
- 다운로드 데이터 포맷
- 원천데이터
    - PCM(WAV) 음성 파일
    - 대상자 및 대화 시나리오 정보를 포함한 음성파일
  - 메타데이터
    - Json 형태
    - 대상자 상세정보 (성별 / 연령 / 지역)
    - 녹음환경 정보 (실내 / 실외 : 대중교통, 거리 등)
    - 대화 주제 및 상세내용

- Json 내 포함 내용
  - dataClass : 데이터 분류
  - colctUnitCode : 수집 방법
  - convrsThema : 대화 주제
  - cityCode : 지역
  - recrdEnvrn : 녹음환경
  - recrdUnit : 녹음도구
  - recorder : 녹음자
  - gender : 성별
  - age : 나이
  - recrdDt : 녹음일시
  - recrdTime : 녹음시간
  - stt : 음성인식 결과
  - fileNm : 파일명
  - recrdQuality : 녹음품질

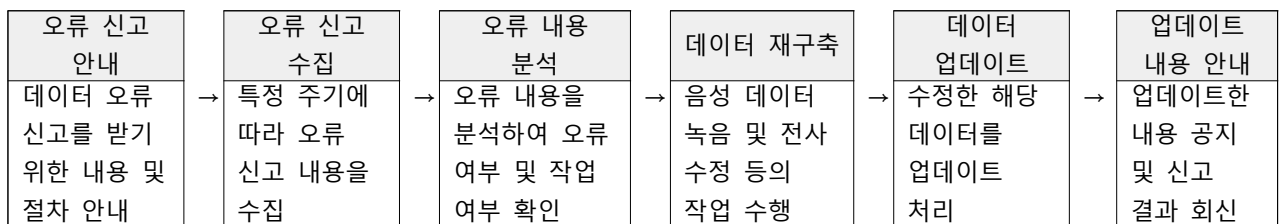
### 2.6.3 데이터 유지보수

#### ○ 데이터 유지보수 목표

- 구축하고 공개한 AI데이터를 사용자 이용 시 오류사항 등이 발견되면, 본 사업 기간 종료 이후에도 유지보수를 지원하고 해당 오류사항을 수정하여 AI데이터를 제공 필요
- 본 사업이 단발성 구축 사업에 그치지 않고, 지속적으로 AI데이터를 제공하고 이용자의 만족도 제고 필요

#### ○ 데이터 유지보수 방안

- 데이터 유지보수 시 오류 신고 안내, 오류 신고 수집, 오류 내용 분석, 데이터 재구축, 데이터 업데이트, 업데이트 내용 안내와 같은 사이클로 유지보수 작업 수행



#### ○ 데이터 오류 신고 안내

- 본 사업을 통해 구축한 AI데이터를 AI 허브 등의 사이트에 공개 시, 유지보수 목적으로 유지보수 담당자의 연락처(전화번호, 이메일 등)를 데이터 정보와 같이 공유하고 데이터에 이상이 있을 경우 사용자가 별도로 고지하도록 유도
- 데이터 오류 시 하기 내용을 포함하여 오류내용을 신고 받고, 수정 이후에 수정결과를 회신
  - 데이터 유형(일반남여, 노인남여, 소아남여/유아 등 혼하, 한국인 외래어)
  - 오류가 있는 파일명
  - 오류 내역(자세히 기재 필요, ex) 전사 오류: 배고파 -> 배고파)
  - 수정 결과를 확인 받을 수 있는 이메일 주소

○ 데이터 오류 신고 수집 및 분석에 따른 재구축

- 주단위 또는 월단위의 특정 기간 동안 데이터 오류 내용을 수집하고, 수집이 완료되면 수집된 내용을 분석
- 데이터 수집 기간은 실제 AI데이터 공개 이후 이용수와 오류내용 수집 건수 등에 따라 한국정보화진흥원과 협의하여 결정
- 오류사항을 분석하여, 음성 오류 내용이 확인되면, 해당 데이터의 유형과 동일한 유형의 구축 작업자를 제외하여 해당 부분 다시 녹음 수행
- 오류사항 중 전사 오류인 경우는 추가 녹음 없이 전사 텍스트 부분만 별도 수정

○ 데이터 업데이트 및 업데이트 내용 안내

- 수정 작업 완료 후 수정한 데이터를 업데이트 하고 수정한 내용 등은 공지사항 등의 게시판을 통해 공개
- 각 오류 신고자에게는 수정한 결과를 별도로 연락하여 업데이트 내용 공유
- 오류 신고 건수 중 분석 결과 오류가 아닌 것으로 판별된 것도 오류 신고자에게 해당 내용에 대해 공유

## 별첨. 데이터 구축 일정

			목표량	확보량	진행률	10						11				12			
						4 (09.21 ~25)	5 (추석) (09.28 ~10.02)	1 (10.05 ~09)	2 (10.12 ~16)	3 (10.19 ~23)	4 (10.26 ~30)	1 (11.02 ~06)	2 (11.09 ~13)	3 (11.16 ~20)	4 (11.23 ~27)	1 (11.30 ~12.04)	2 (12.07 ~11)	3 (12.14 ~18)	4 (12.21 ~25)
						목표	목표	목표	목표	목표	목표	목표	목표	목표	목표	목표	목표	목표	목표
모집명 (명)	일반	셀바스	200	200	100.0%	10		65	75	50									
		다이퀘스트 (크라우드웍스)	800	279	34.9%	100		100	100	100	100	100	100	100					
		잉글리시헌트	1,000	0	0.0%						100	200	200	200	200	100			
	노인	셀바스	100	0	0.0%									30	30	40			
		다이퀘스트	100	0	0.0%							25	25	25	25				
		원더풀플랫폼	800	300	37.5%				50	100	120	120	120	120	120	50			
	소아	셀바스	100	20	20.0%							30	30	40					
		다이퀘스트	100	100	100.0%							25	25	25	25				
		잉글리시헌트	800	100	12.5%					50	100	200	200	200	50				
	외래어	셀바스	200	200	100.0%				50	50	50	50							
		다이퀘스트 (크라우드웍스)	800	747	93.4%	100		100	100	100	100	100	100	100					
		잉글리시헌트	1,000	800	80.0%					100	200	200	200	200	100				
	소계		6,000	2,746	45.8%	210	0	265	375	550	770	1,050	1,000	1,040	550	190	0	0	0
수집 (시간)	일반	셀바스	400	400	100.0%	20		120	150	110									
		크라우드웍스	1,600	483	30.2%	200		200	200	200	200	200	200	200					
		잉글리시헌트	2,000	0	0.0%						200	400	400	400	400	200			
	노인	셀바스	300	0	0.0%											15	85	100	100
		다이퀘스트	300	0	0.0%							75	75	75	75				
		원더풀플랫폼	2,400	120	5.0%				10		200	450	500	500	390	350			
	소아	셀바스	300	0	0.0%							15	85	100	100				
		다이퀘스트	300	0	0.0%							75	75	75	75				
		잉글리시헌트	2,400	50	2.1%					150	300	600	600	600	150				
	외래어	셀바스	400	328	82.0%					20	150	150	80						
		크라우드웍스	1,600	833	52.1%	200		200	200	200	200	200	200	200					
		잉글리시헌트	2,000	220	11.0%					200	400	400	400	400	200				
	소계		14,000	2,434	17.4%	420	0	520	560	880	1,650	2,565	2,615	2,550	1,390	565	85	100	100
가공 (시간)	일반	셀바스	400	400	100.0%			30	90	150	130								
		크라우드웍스	1,600	0	0.0%						200	200	200	200	200	200	200	200	
		다이퀘스트	2,000	0	0.0%							200	400	400	400	400	200		
	노인	셀바스	300	0	0.0%												50	150	100
		다이퀘스트	300	0	0.0%								75	75	75	75			
		원더풀플랫폼	2,400	0	0.0%								300	600	600	600	300		
	소아	셀바스	300	0	0.0%									15	105	150	30		
		다이퀘스트	300	0	0.0%								75	75	75	75			
		잉글리시헌트	2,400	0	0.0%						150	300	600	600	600	150			
	소계																		

	외 래 어	셀바스	400	30	7.5 %						30	90	150	130					
		크라우드웍스	1,600	0	0.0 %						200	200	200	200	200	200	200	200	
		앙글리시현트	2,000	0	0.0 %						200	400	400	400	400	200			
		소계	14,000	430	3.1%	0	0	30	90	150	910	1,390	2,400	2,695	2,655	2,050	980	550	100
건 중 간 (시간)	인 반	아임클라우드	4,000	0	0.0 %								500	700	700	700	700	500	200
	노 인	아임클라우드	3,000	0	0.0 %								600	675	675	675	75	50	250
	소 아	아임클라우드	3,000	0	0.0 %								300	675	690	780	375	80	100
	외 래 어	아임클라우드	4,000	0	0.0 %								690	750	730	600	500	430	300
		소계	14,000	0	0.0%				0	0	0	0	2,090	2,800	2,795	2,755	1,650	1,060	850