



21년도 인공지능 학습용 데이터 구축 가이드라인

< 외국인 발화 한국어 음성 데이터 >

인공지능 데이터 구축	사업 총괄	CSLEE®
	데이터 설계	CSLEE®  이화여자대학교  세종대학교
	데이터 수집 및 정제	CSLEE® Dream BIT  이화여자대학교  세종대학교
	데이터 가공	CSLEE® Dream BIT  DGRAIB 데이터로 미래를 만든다 디그립
	데이터 검수	CSLEE® Dream BIT
	클라우드 소싱	CSLEE® Dream BIT  이화여자대학교  세종대학교
	저작도구 개발	CSLEE®
	AI모델 개발	 ActionPower
가이드라인 작성	씨에스리	오상아
가이드라인 버전	ver 1.0 ('22. 1. 28)	

목 차

1. 데이터 명세 정보	1
1.1 데이터 정보 요약	1
1.2 데이터 포맷	1
1.3 어노테이션 포맷	2
1.4 데이터 구성	3
1.5 데이터 통계	4
1.6 원시데이터 특성	5
1.7 기타 정보	5
 2. 데이터 구축 가이드	 6
2.1 데이터 구축 개요	6
2.2 문제정의	6
2.3 수집·정제	6
2.4 어노테이션/라벨링	7
2.5 검수	10
2.6 활용	10

1. 데이터 명세 정보

1.1 데이터 정보 요약

데이터 이름	외국인 발화 한국어 음성 데이터	
활용 분야	외국인이 발화하는 한국어 음성 인식 등 음성 및 자연어 처리 모델	
데이터 요약	주어진 대본을 읽거나, 주어진 질문에 자유롭게 한국어로 응답한 외국인의 음성 데이터	
데이터 출처	자체적으로 수집한 외국인 음성	
데이터 이력	배포버전	ver 1.0
	개정이력	신규
	작성자/ 배포자	오상아 / 오상아

1.2 데이터 포맷

- 데이터는 음성, json, csv의 쌍으로 구성
- 음성 오디오 파일은 불특정 다수 AI 학습모델들에서의 보편적 활용을 고려하여 wav 파일로 구축
- 메타데이터는 AI 툴에서의 활용성, 유연성, 확장성, 상호운용성, 처리성능 등을 고려하여 json과 csv 파일을 병용해 구축
- 하나의 json 데이터는 녹음 음성 정보, 전사 결과, 녹음자 정보 등의 세부 정보를 포함
- csv 데이터는 json 데이터와 같은 정보를 포함하며 file_info, transcription, basic_info, residence_info, skill_info 등 json 데이터에서의 상위 depth는 별도로 작성하지 않음

0:00 / 0:19

🔊

⋮

fileName	speakerID	sentenceID	recordUnit	recordQuality	recordDate	recordTime	Reading	ReadingLabelText	Question	AnswerLabelText	SentenceSpeechLV	SpeakerID	gender	birthYear	eduBackground	Country	ResidencePeriod	ResidenceCity	LanguageClass	MotherTongue	SelfAssessment	TopicGrade	LearningPeriod	LearningSource
VN10QC226_VN0001_20210809.wav	VN0001	VN10QC226	ios	16bit 16kHz MONO	2021-08-09 22:44	9.384			당신은 어느 나라에서 왔나요? 영어로 한국 남자 한국어 문 이해 하 경험해서 한국어는 무엇인가요? 왜 오게 됐어요	중	VN0001	F	1995	고졸	VN	5년 이상	KR-41	베트남어	베트남어	중	4	36	기타	

JSON

오디오 파일

전사 결과

녹음자 정보

```

{
  "fileName": "VN11RC001_VN0001_20211124.wav",
  "file_info": {
    "speakerID": "VN0001",
    "sentenceID": "VN11RC001",
    "recordUnit": "ios",
    "recordQuality": "16bit 16kHz MONO",
    "recordDate": "2021-11-24 12:01:01",
    "recordTime": "9.316"
  },
  "transcription": {
    "Reading": "잠시만, 내 지갑이랑 열쇠가 어디 갔지? 유람아, 미안한데 나 물건이 없어진 거 같아.",
    "ReadingLabelText": "잠시만 내 지갑이랑 열쇠가 어디 갔지 유람아 미안한데 나 물건이 없어진 거 같아",
    "Question": "",
    "AnswerLabelText": "",
    "SentenceSpeechLV": "상"
  },
  "SpeakerID": "VN0001",
  "basic_info": {
    "gender": "F",
    "birthYear": "1995",
    "eduBackground": "고졸"
  },
  "residence_info": {
    "country": "VN",
    "residencePeriod": "5년 이상",
    "residenceCity": "KR-41"
  },
  "skill_info": {
    "languageClass": "베트남어",
    "motherTongue": "베트남어",
    "selfAssessment": "중",
    "topicGrade": "4",
    "LearningPeriod": "36",
    "learningSource": "기타"
  }
}

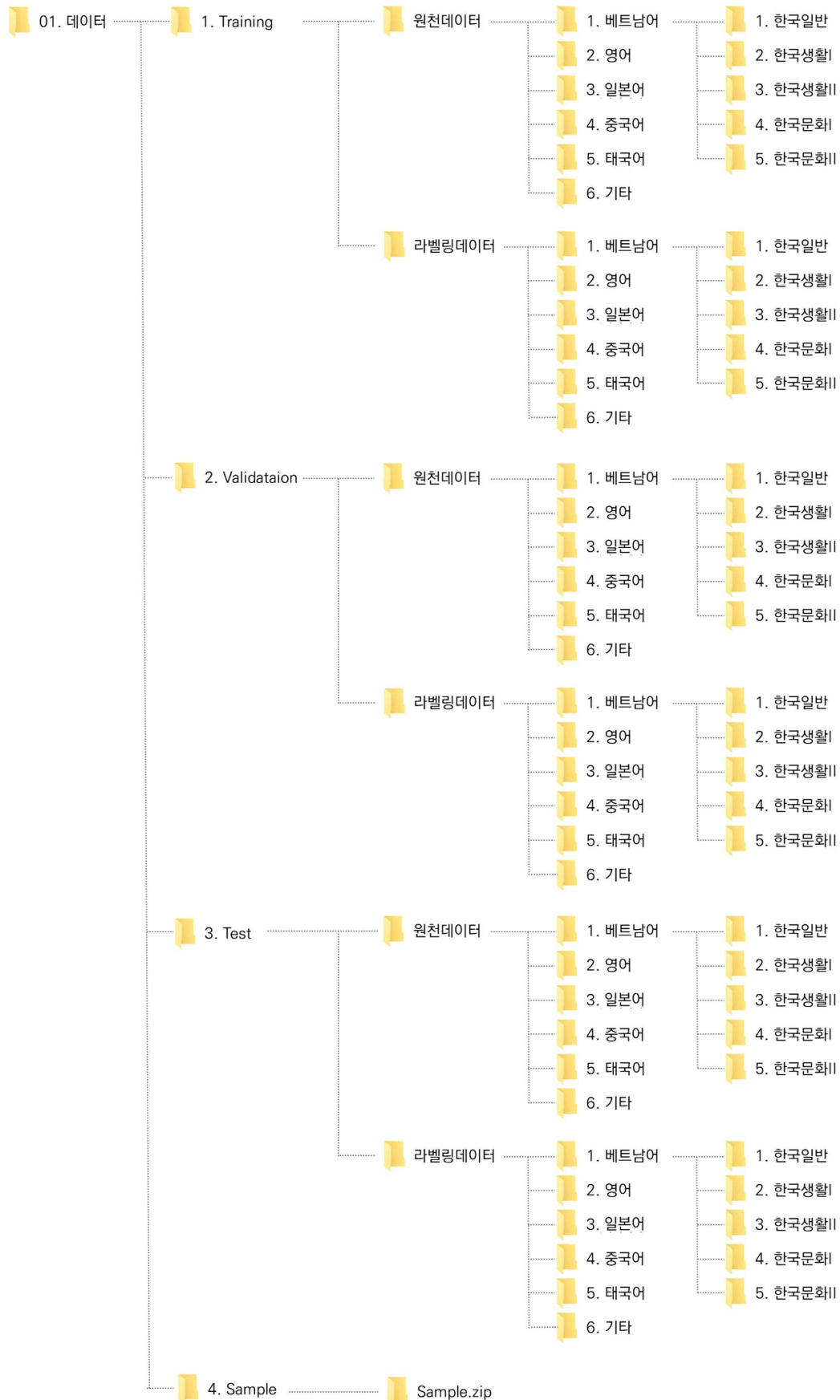
```

1.3 어노테이션 포맷

NO	항목		타입	길이		단위	필수여부	유효값
	항목명	항목 설명		최소	최대			
1	fileName	파일명	string	29	29	-	Y	-
2	file_info	녹음파일 정보	Object	-	-	-	Y	-
2-1	speakerID	녹음자ID	string	6	6	-	Y	-
2-2	sentenceID	문장ID	string	9	9	-	Y	-
2-3	recordUnit	녹음단말기 유형	string	3	7	-	Y	android, ios
2-4	recordQuality	녹음 품질	string	16	16	-	Y	16bit 16KHz MONO
2-5	recordDate	녹음날짜	string	19	19	-	Y	-
2-6	recordTime	녹음 시간	string	1	7	초	Y	-
3	transcription	대본/질문 정보	Object	-	-	-	Y	-
3-1	Reading	대본 내용	string	0	100	-	N	-
3-2	ReadingLabelText	대본읽기 전사 결과	string	0	200	-	N	-
3-3	Question	질문 내용	string	0	150	-	N	-
3-4	AnswerLabelText	질문답변 전사 결과	string	0	1300	-	N	-
3-5	SentenceSpeechLV	문장 구사 실력	string	1	1	-	Y	상, 중, 하
4	SpeakerID	녹음자ID	string	6	6	-	Y	-
5	basic_info	녹음자 기본 정보	Object	-	-	-	Y	-
5-1	gender	성별	string	1	1	-	Y	M, F
5-2	birthYear	출생연도	string	4	4	년	Y	-
5-3	eduBackground	최종학력	string	2	4	-	Y	초졸, 중졸, 고졸, 대졸, 석사이상
6	residence_info	녹음자 거주 정보	Object	-	-	-	Y	-
6-1	country	국적	string	2	2	-	Y	-
6-2	residencePeriod	국내 체류 기간	string	5	11	-	Y	1년 미만, 1년 이상 3년 미만, 3년 이상 5년 미만, 5년 이상
6-3	residenceCity	거주 지역	string	4	6	-	Y	-
7	skill_info	녹음자 실력 정보	Object	-	-	-	Y	-
7-1	languageClass	언어 분류	string	2	4	-	Y	베트남어, 영어, 일본어, 중국어, 태국어, 기타
7-2	motherTongue	모국어	string	2	9	-	Y	-
7-3	selfAssessment	자가평가실력	string	1	1	-	Y	상, 중, 하
7-4	topikGrade	TOPIK 등급	string	1	5	-	Y	1, 2, 3, 4, 5, 6, 해당없음
7-5	LearningPeriod	한국어 학습 기간	string	1	3	개월	Y	-
7-6	learningSource	한국어 학습 방법	string	2	13	-	Y	학교수업(대학 포함), 학원, 인터넷 강의, 서적, 한국인 친구, 영화/음악, 기타

1.4 데이터 구성

- 본래 폴더 구조



- 오브젝트 스토리지 최종 데이터 업로드 시 폴더 구조



1.5 데이터 통계

1.5.1 데이터 구축 규모

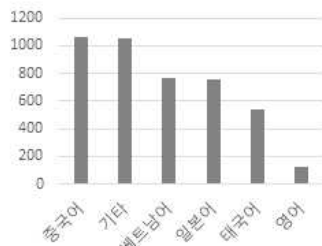
- 외국인 발화 한국어 음성 약 4,302시간 (16bit 16kHz MONO, 1,159,665건, 461.66 GB)
- 어노테이션 json 파일 1,159,665건, 어노테이션 csv 파일 1,159,665건 (약 1.99 GB)

1.5.2 데이터 분포

언어분류	세트	원천데이터	라벨링데이터			총합	녹음 시간		
		wav 건수	csv 건수	json 건수	합계		녹음 시간(s)	녹음 시간 (h)	총합(h)
베트남어	한국일반	38,877	38,877	38,877	77,754	388,164	515,765	143	765
	한국생활1	38,597	38,597	38,597	77,194		538,629	150	
	한국생활2	40,990	40,990	40,990	81,980		605,280	168	
	한국문화1	39,948	39,948	39,948	79,896		565,390	157	
	한국문화2	35,670	35,670	35,670	71,340		528,701	147	
영어	한국일반	17,316	17,316	17,316	34,632	63,108	240,206	67	124
	한국생활1	3,395	3,395	3,395	6,790		45,255	13	
	한국생활2	5,008	5,008	5,008	10,016		75,029	21	
	한국문화1	3,434	3,434	3,434	6,868		49,452	14	
	한국문화2	2,401	2,401	2,401	4,802		36,110	10	
일본어	한국일반	39,849	39,849	39,849	79,698	433,804	472,147	131	760
	한국생활1	40,153	40,153	40,153	80,306		496,130	138	
	한국생활2	41,905	41,905	41,905	83,810		545,035	151	
	한국문화1	47,516	47,516	47,516	95,032		592,925	165	
	한국문화2	47,479	47,479	47,479	94,958		630,041	175	
중국어	한국일반	64,905	64,905	64,905	129,810	585,186	802,864	223	1,065
	한국생활1	57,066	57,066	57,066	114,132		728,725	202	
	한국생활2	53,755	53,755	53,755	107,510		725,622	202	
	한국문화1	63,020	63,020	63,020	126,040		824,853	229	
	한국문화2	53,847	53,847	53,847	107,694		750,708	209	
태국어	한국일반	35,639	35,639	35,639	71,278	279,384	477,425	133	538
	한국생활1	28,595	28,595	28,595	57,190		387,965	108	
	한국생활2	30,115	30,115	30,115	60,230		433,986	121	
	한국문화1	25,099	25,099	25,099	50,198		345,354	96	
	한국문화2	20,244	20,244	20,244	40,488		291,425	81	
기타	한국일반	70,781	70,781	70,781	141,562	569,684	903,403	251	1,051
	한국생활1	62,371	62,371	62,371	124,742		815,647	227	
	한국생활2	61,079	61,079	61,079	122,158		839,807	233	
	한국문화1	45,313	45,313	45,313	90,626		600,919	167	
	한국문화2	45,298	45,298	45,298	90,596		623,819	173	
총합		1,159,665	1,159,665	1,159,665	2,319,330	2,319,330	5,488,618	4,302	4,302

합계 : 녹음 시간 (h)

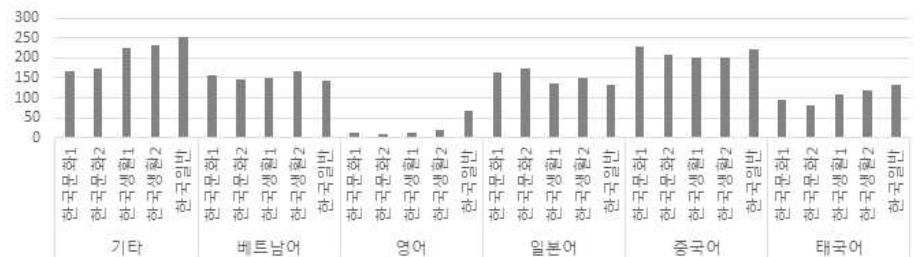
언어분류별 녹음시간



언어분류 ▼

합계 : 녹음 시간 (h)

언어분류-세트 별 녹음시간



언어분류 ▼ 세트 ▼

+ -

1.5.3 기타 활용 통계

- 관련 연구를 통해 도출한 통계 정보 없음

1.6 원시데이터 특성

1.6.1 대상분류

- 원시데이터 수집 대상은 국내 및 국외에 거주 중인 외국인의 실제 한국어 발화 음성이다.
- 이에 원시데이터의 대상 분류는 '실제'로 구분된다.

1.6.2 제약조건

- 원시데이터 수집 조건은 일부 제약 있음(semi-constrained)으로 구분한다.
- 본 사업은 '대본읽기'와 '질문에 답변하기' 두 가지 유형으로 나누어 외국인 한국어 발화 음성을 수집하였다. '대본읽기' 유형의 경우 사전에 주어진 대본 내용이 있으므로 '제약있음'에 해당한다. 반면 '질문에 답변하기' 유형은 주어진 질문에 대해 작업자가 자유롭게 답변한 음성을 수집하므로 '제약없음'에 해당한다.

1.6.3 속성

- 원시데이터의 정제 이후 일반 속성으로는 음성 시작 시 휴지 구간 0.3초와 16bit 16kHz MONO 품질이 있다.

1.7 기타정보

1.7.1 포괄성

원시데이터는 외국인이 발화한 한국어 음성으로 언어분류는 크게 베트남어, 영어, 일본어, 중국어, 태국어, 기타 6가지로 구분하여 수집하였다. 6가지 언어분류는 국내 한국어 학습자, 국외 한국어 학습자, 국내 체류 외국인 (결혼이민자, 외국인 근로자, 대학교 어학당 유학생) 등의 국적 분포 조사에 근거하여 공통적으로 높은 분포를 보인 언어분류로 선정하였다. 이에 따라 본 사업에서 수집한 원시데이터는 국내 및 국외에서 한국어 사용 및 학습 빈도가 높은 외국인의 한국어 발화 음성과 언어분류 별로 원어민과 달리 특징적으로 발생하는 모국어 부정전이 현상을 모두 포괄한다.

1.7.2 독립성

데이터는 녹음자의 음성과 개인정보를 포함하고 있으나, 원시데이터 수집 시 녹음 작업자에게 녹음 음성을 포함해 데이터로 공개될 개인정보를 안내하고 개인정보 수집 및 이용 동의를 받았다. 또한, 해당 내용에 동의하지 않은 경우 녹음 작업에 참여할 수 없도록 하였으며 녹음 데이터를 검수하여 민감정보와 개인을 특정할 수 있는 개인정보가 포함된 음성은 제거하도록 하였다. 이에 데이터와 관련해 법률 개정 및 민감정보 등 의존하고 있는 사항은 없다.

1.7.3 유의사항

- 유의사항 없음

1.7.4 관련 연구

- 관련 연구 없음

1.8 데이터 구축 개요

구축 단계	세부 절차	설명
1. 수집	1-1. 대본 및 질문 내용 구성	원시데이터 수집의 기반이 되는 대본 및 질문 내용 선정 및 구성
	1-2. 녹음 음성 수집	주어진 대본 내용을 읽는 외국인 음성 데이터 수집
		주어진 질문에 대해 자유롭게 답변하는 외국인 음성 데이터 수집
2. 정제	2-1. 부적합 데이터 반려 및 삭제	음성 데이터 수집 기준에 부적합한 음성 데이터 재녹음하도록 반려 및 재검수
	2-2. 데이터 정제	음성 데이터 품질 통일 작업 진행
		음성 데이터 휴지 구간 일괄부여 작업 진행
3. 가공	3-1. 녹음 음성 전사	녹음 음성당 전사 작업자 2명씩 배정하여 전사 규칙에 따라 전사 진행
4. 검수	4-1. 녹음 음성 전사 결과 비교	녹음 음성별 전사 결과 2건 동일 여부 확인
	4-2. 녹음 음성 전사 불가 건 제거	녹음 음성 전사 과정 중 전사 불가 기준에 해당하는 것으로 리포트 된 음성 제거
	4-3. 전사 결과 상이 건 3차 전사	녹음 음성별 전사 결과 2건 상이한 건 3차 전사 작업 진행
	4-4. 녹음 음성별 전사 결과 선별	녹음 음성별 전사 결과 2건이 동일한 경우, 2건 중 하나를 선별
		녹음 음성별 전사 결과 2건이 상이한 경우 3번째 전사 작업 결과를 선별

1.9 문제정의

1.9.1 임무 정의

법무부 통계에 의하면, 코로나 전염병 발생 이전인 2019년 기준으로 대한민국에 체류하고 있는 외국인의 수는 약 250만 명으로 전체 인구의 4.87%에 달하며, 매년 증가하는 추세를 보이고 있다. 또한, 국립국제교육원 통계에 따르면 전 세계 80여국에서 한국어능력시험 TOPIK에 응시한 수험자 수는 2015년 17만여 명에서 2019년 35만여 명으로 5년 사이 두 배 이상 증가하였다. 취업, 유학 및 기타 목적으로 한국어를 배우는 인구가 급속도로 증가하는 경향을 보이는 것이다.

최근 K-POP의 세계적인 인기, 한국영화와 드라마의 인기 상승 등으로 한국 문화의 영향력이 폭발적으로 증가하고 있으며, 코로나 전염병 위기에 대한 대한민국 정부와 국민의 성공적인 대처로 인해 국제 사회에서의 한국의 위상 제고와 인식의 개선이 이루어지고 있다. 이에 앞으로도 한국과 한국 문화에 관한 관심은 더욱 증가할 것으로 보이며, 자연스럽게 한국어 학습과 한국 방문 및 체류 인구의 증가로 이어질 것으로 예상된다. 이에 따라 앞으로 외국인 화자의 한국어 음성인식에 대한 수요가 큰 폭으로 증가할 것으로 예상되며, 이에 따라 외국인 음성을 인식하는 인공지능의 연구와 개발이 절실히 요구되는 상황이다.

외국인 화자는 모국어의 음성적, 음운적 특징이 한국어와 다르므로 한국어 발음을 정확하게 구사하기 힘들다. 이에 외국인 화자의 한국어 발음은 한국인 모국어 화자의 발음과 상당히 다르다. 이러한 발음적 특성은 쉽게 고쳐지는 것이 아니므로, 한국어 초급 학습자는 물론이고 오랜 기간 한국에 거주하여 고급 수준의 한국어를 구사하는 화자에게도 한국어 원어민과는 상이한 발음이 나타난다. 이에 한국인 모국어 화자의 음성발화 데이터에 기반하여 학습한 인공지능은 외국인 화자의 상이한 발화를 인식하기 어려우며, 음성 인식률이 현저히 떨어질 수밖에 없다. 한국을 방문하거나 체류하는 외국인은 한국어를 구사하더라도 발음의 한계로 인하여 내국인을 대상으로 개발된 각종 전화 상담 및 정보 제공 서비스, 주문/배송조회 확인 서비스, 119 등 응급 전화 서비스 등을 이용하기 어려우며, 공적, 사적 서비스 영역에서 외국인을 위한 콜센터 및 음성 서비스는 부족한 상황이다.

현재 외국어 화자의 음성 데이터는 국내에서 제대로 구축된 바가 없어 이에 관한 연구 또한 매우 부족한 상황이다. 데이터의 부족으로 정확한 인식률 차이를 알 수는 없으나, 현장에서는 외국인 화자의 인식률이 모국어 화자 인식률과 비교하면 대략 30% 이상 떨어질 것으로 짐작하고 있다(ETRI 음성 인식팀). 이에 외국인 화자의 한국어 음성발화 데이터 구축과 이를 기반으로 한 인공지능의 학습연구 및 개발이 시급히 진행되어야 한다.

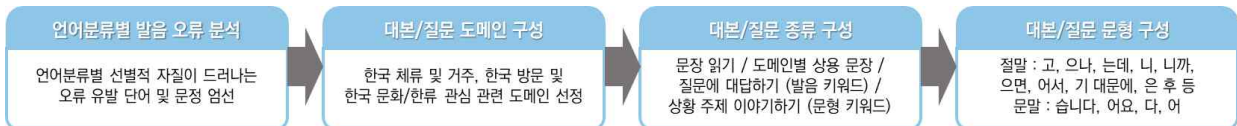
1.9.2 데이터 구축 유의사항

데이터는 녹음자의 음성과 개인정보(국적, 모국어, 출생연도, 성별, 한국어 학습 기간 등)를 포함하고 있으므로 개인정보 수집 및 이용 관련 법적 문제가 있을 수 있으나, 데이터 수집 시 참여한 모든 녹음 작업자에게 음성을 포함한 개인정보 수집 및 이용 동의서를 사전에 획득하고 AI Hub 공개를 고지하여 데이터 제공으로 인해 발생할 수 있는 법률적 문제를 방지 및 해결하였다. 또한, 대본과 질문 구성 시 성차별, 정치적 성향, 종교 등 사회적 민감정보와 관련한 내용은 포함되지 않도록 하였다. 대본 읽기 유형 녹음의 경우 사전에 작성된 대본을 중심으로 녹음이 이루어지므로 개인정보나 민감정보를 포함하지 않는다. 또한, 질문에 답변하기 유형의 경우 녹음 이전에 민감정보에 대해 주의하여 답변할 수 있도록 크라우드소싱 인력에 사전 고지하였으며 녹음 음성에서 개인을 특정할 수 있는 개인정보나 민감정보가 포함되어 있을 경우 검수 과정에서 제외 처리하여 개인정보 유출 소지를 해결하였다.

1.10 수집·정제

1.10.1 원시데이터 선정

원시데이터는 주어진 대본을 그대로 읽거나 주어진 질문에 자유롭게 응답한 외국인 발화 한국어 음성이다. 원시데이터 수집 시 외국인 발화자의 모국어에 따라 다르게 나타나는 선별적 음성 자질이 잘 드러날 수 있도록 음성 데이터를 구축하여야 할 필요가 있다. 또한, 한정된 데이터양을 고려할 때 한국어를 사용하는 외국인의 방문 및 체류 목적에 맞는 특화된 도메인으로 내용을 집중하여 음성 데이터를 구축할 필요가 있다. 특화된 도메인에서 빈번히 사용되는 단어와 문장 패턴으로 이루어진 데이터를 확보한다면 인공지능이 활용되는 단계에서 음성 인식을 높이고 관련 서비스에서의 활용 가능성을 높일 수 있기 때문이다. 또한, 한국어의 화용 상황별 문형 및 문미의 다양성을 고려하여 도메인별, 상황별로 다양한 종류의 문형으로 된 녹음 음성의 확보가 필요하다. 이에 원시데이터의 기본 베이스가 되는 대본 및 질문 내용 구성 시 외국인의 모국어별 오류 패턴, 외국인 방문 및 체류 목적에 맞는 특화된 도메인 선정, 도메인 및 상황별 다양한 문형 선택을 고려하여 구성할 수 있도록 하였다.



이러한 과정에 따라 구성된 세트는 한국일반, 한국생활I, 한국생활II, 한국문화I, 한국문화II로 총 5개이다. 각 세트에는 질문 유형을 포함하여 각각 10개씩 스크립트 분류를 선정하여 구성하였다. 해당 스크립트에는 '기차표 예매', '배달음식주문', '영화 예매', '지하철타는법', '분실물찾기'처럼 국내 체류 및 거주 외국인이 실생활에서 익숙하게 접할 수 있는 상황과 관련된 내용과 '결혼호칭'과 '한국의예절문화'와 같이 국내 체류 및 거주 시 숙지할 필요가 있는 한국의 호칭 및 예절문화와 관련된 내용이 포함되어 있다. 또한, 한국 방문 및 한국 문화/한류 관심과 관련하여 '서울의관광명소', '한국의축제', '한국의아름다움', '부산여행'처럼 한국 관광과 관련된 내용과 '한국드라마', '한국영화', '한국어학습동기(케이팝)'과 같이 한국 문화와 관련된 내용을 포함하도록 하였다. 스크립트별로 대본은 25개씩으로 구성하였으며, 질문 유형은 75개 질문으로 구성하였다. 이에 따라 세트별 총 대본 및 질문 개수는 300개이다.

세트	한국일반	한국생활I	한국생활II	한국문화I	한국문화II
스크립트	기차표예매	결혼호칭	결혼식	경복궁관람	비행기표예매
	날씨	다문화	빨리빨리	공항수속상황	서울의관광명소
	모임	대중교통	쌈문화	선물고르기	식당예약
	배달음식주문	병문안	육아휴직	영화예매	한국어학습동기(케이팝)
	분실물찾기	부산여행	제주도	일상의한국문화	한국영화
	선물	생활패턴	추석	지하철타는법	한국의여행지
	소비	설	택배알바	취미생활	한국의예절문화
	약처방	여행지	한국식단	한국드라마	한국의축제
	한국음식만들기	치킨요리	한의원	한국어수강신청	한국의아름다움
	질문	질문	질문	질문	질문

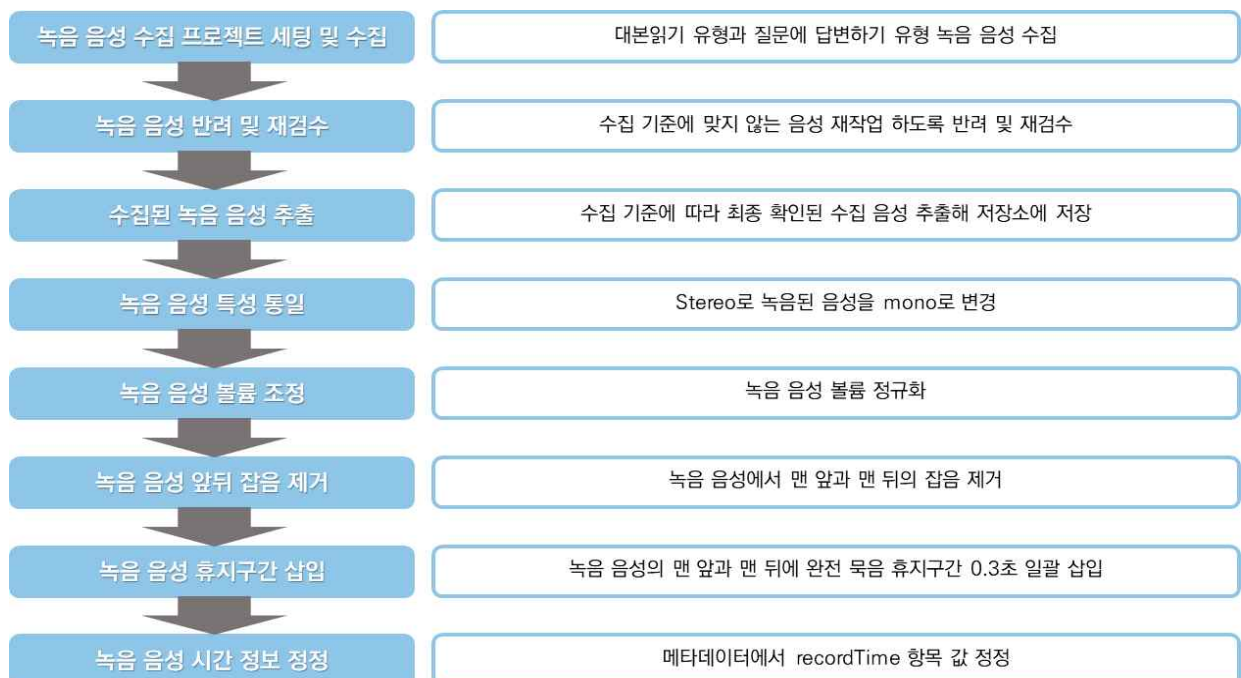
추후 녹음 음성을 문장 단위로 자르거나 이어 붙이는 등의 작업이 정제 단계에서 이루어질 경우, 오랜 시간이 소요되고 음성이 훼손되는 문제가 발생할 수 있다. 이에 따라 각 대본 내용은 한국어 원어민에 비해 다소 느린 외국인의 발화 속도를 고려하여 약 12초 내외로 녹음할 수 있는 분량으로 구성하였다. 또한, '질문에 답변하기' 유형에서도 약 10초에서 30초 이내로 녹음하도록 안내하였다. 이에 따라 각 녹음 음성은 자동으로 약 12초 분량으로 나누어 녹음되어 추후 정제 작업에서 녹음 음성을 자르거나 이어 붙이는 등의 작업으로 인해 발생할 수 있는 문제를 방지하였다.

외국인 녹음 작업자는 이처럼 구성된 대본 및 질문 내용을 바탕으로, '대본 읽기' 유형은 주어진 대본 내용을 그대로 읽는 음성을 녹음하도록 하고 '질문에 답변하기' 유형은 주어진 질문 내용을 바탕으로 자유롭게 답변하는 음성을 녹음하였다. 이에 따라 작업자별로 전체 세트를 녹음하였을 경우, 300개 이상의 약 12초 분량의 녹음 음성을 원시데이터로 확보할 수 있다.

원시데이터 획득 후 정제, 라벨링, 검수 과정에서 기준 미충족으로 버려지는 데이터양을 고려하여 구축 목표 4,000시간의 1.1배인 4,400시간을 수집하는 것을 목표로 하였다. 외국인의 한국어 발음 속도에 따라 녹음 1건당 speech signal 이 전사된 시간은 약 12초를 기준으로 하였으며, 이에 따라 목표 녹음 시간과 문장 수를 산정하였다. 산정한 결과는 다음과 같다.

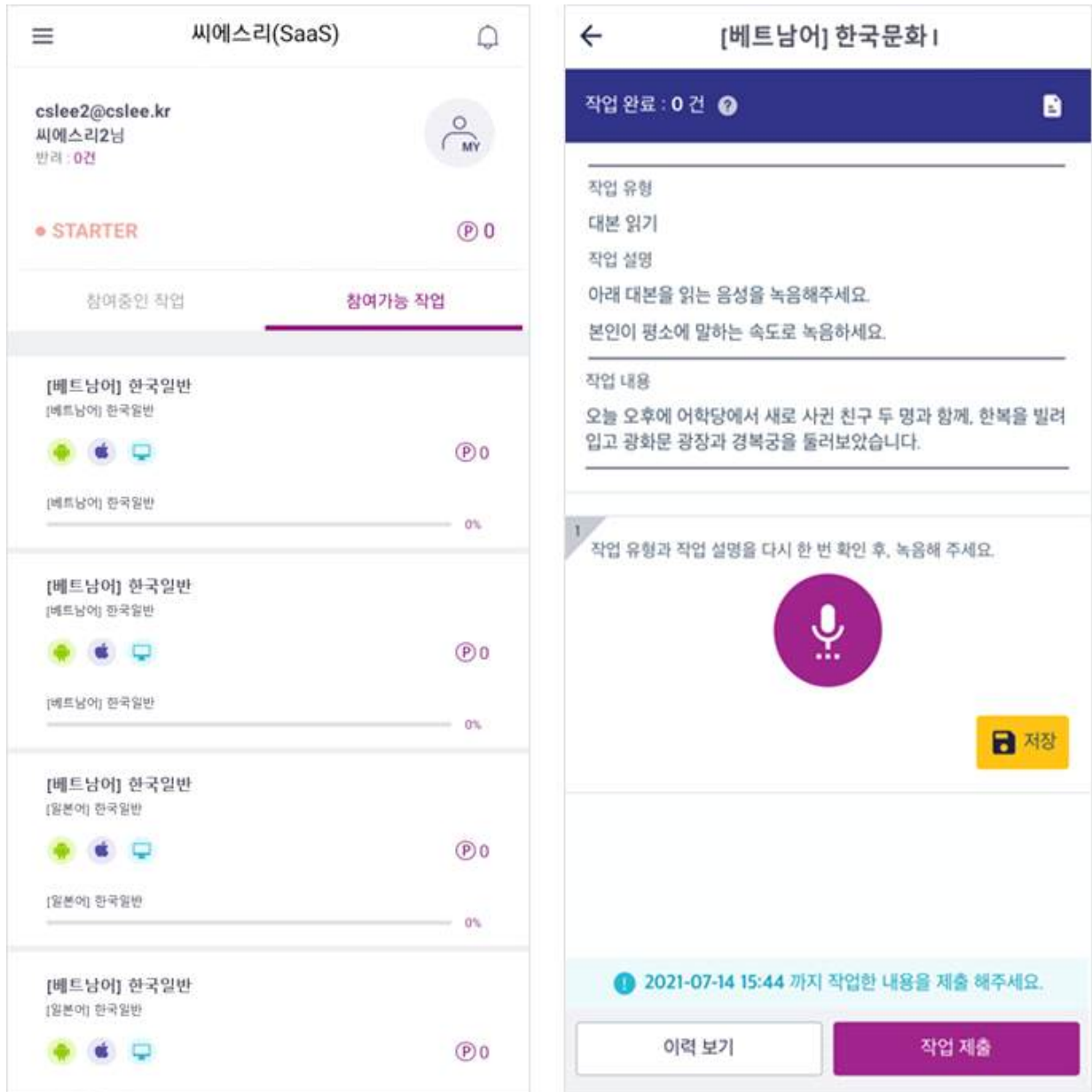
언어분류	구축 목표 시간 (h)	수집 목표 시간 (h)	수집 목표 시간 (s)	수집 목표 문장 수 (문장당 12초 기준)
베트남어	700	770	2,772,000	231,000
영어	100	110	396,000	33,000
일본어	700	770	2,772,000	231,000
중국어	1,000	1100	3,960,000	330,000
태국어	500	550	1,980,000	165,000
기타	1,000	1100	3,960,000	330,000
합계	4,000	4400	15,840,000	1,320,000

1.10.2 수집·정제 절차



- 녹음 음성 수집 프로젝트 세팅 및 수집

원시데이터를 선정한 후, 모든 녹음 음성은 애플리케이션을 통해 수집하도록 한다. 녹음 작업자는 각자 등록된 계정으로 녹음 화면에 접속할 수 있다. 접속 시 화면의 참여 가능 작업 탭에서 보이는 프로젝트는 언어분류와 세트가 조합된 단위로 구성된 프로젝트이다. 각 녹음 작업자는 해당 탭에서 자신의 언어분류에 맞는 프로젝트만 보인다. 프로젝트 탭을 눌러 녹음 작업 화면에 접속할 수 있으며, 녹음 화면에서는 작업유형과 녹음 대상인 대본 혹은 질문 내용을 확인하고 녹음할 수 있도록 구성하였다. 화면에서 녹음 후 제출된 음성과 녹음날짜, 녹음 음성 시간, 녹음 단말기 등 녹음 음성 정보는 제출과 동시에 DB에 저장된다. 이에 따라 녹음 음성 정보에 대해 녹음 작업자는 별도 입력 작업을 하지 않는다.



- 녹음 음성 반려 및 재검수

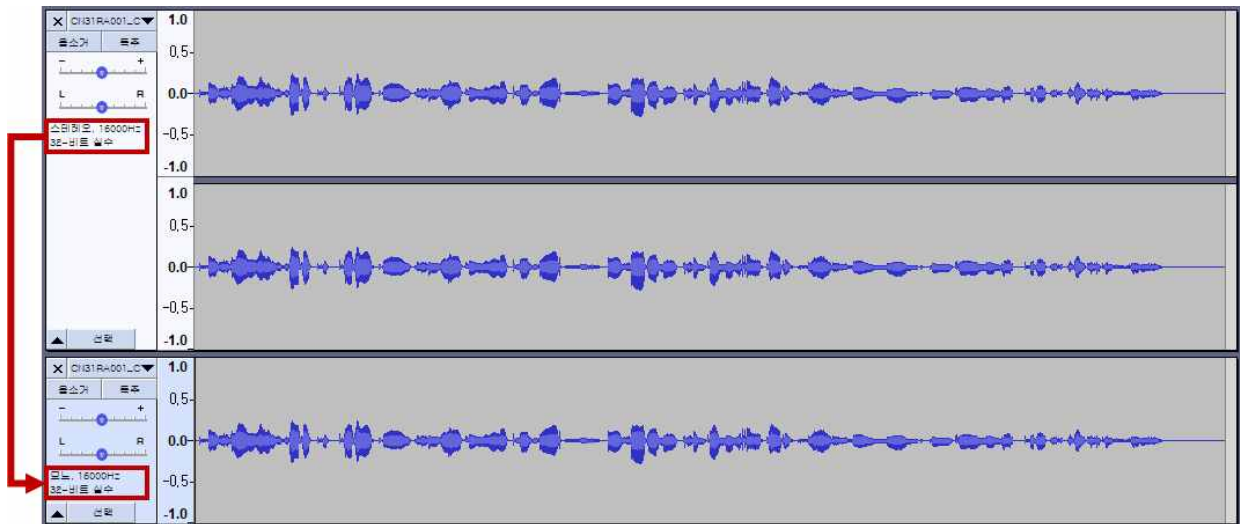
제출된 녹음 음성은 모니터링 페이지에서 확인 및 검수 작업을 진행할 수 있다. 녹음 음성을 듣고, 작업유형에 맞지 않는 작업이거나 자신의 목소리가 아닌 노래를 녹음하는 등 명백하게 불성실한 작업을 하는 등 녹음 가이드라인과 맞지 않는 작업에 대해 해당 작업자가 다시 작업할 수 있도록 반려하였다. 반려 건에 대해 작업자가 재 작업을 하면 해당 작업에 대해 재검수를 진행하였다. 이에 따라 정제 작업 이전에 작업 가이드라인에 맞지 않는 작업 발생을 줄일 수 있다.

- 수집된 녹음 음성 추출

검수 완료된 녹음 음성은 정해진 형식에 따라 음성 데이터는 wav로, 메타데이터는 json과 csv 형식으로 추출하였다. 해당 메타데이터는 아직 전사 단계를 거치지 않았기 때문에 녹음 음성 정보만 포함한다.

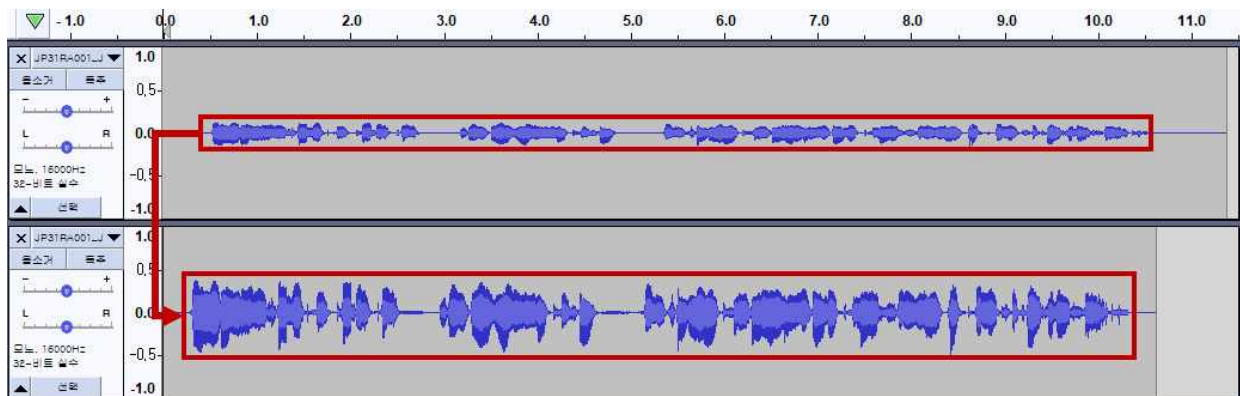
- 녹음 음성 특성 통일

녹음 음성의 품질은 16bit 16kHz MONO로 통일하는 것을 원칙으로 하였으나, 몇몇 음성이 MONO가 아닌 STEREO로 녹음된 것으로 분석되었다. 이에 python 코드를 이용하여 STEREO로 녹음된 음성을 MONO로 변환하는 작업을 진행하였다. 또한, 무료 음성 분석 툴인 audacity를 이용하여 각 음성의 품질이 MONO로 잘 변환되었는지 재차 확인하였다.



- 녹음 음성 볼륨 조정

녹음 작업 시 각 녹음 작업자는 각기 다른 조용한 환경에서 자신의 스마트폰을 이용해 작업하므로 각 음성 데이터마다 파형의 크기가 다르다. 이에 배경음과 발화 음성을 파형으로 구분하여 잘라내는 작업을 진행할 때 하나의 기준으로 배경음과 발화를 구분할 경우, 발화 음성이 배경음과 함께 잘려나갈 위험이 있다. 이러한 점을 고려하여 녹음 음성의 앞뒤 잡음 제거 작업 이전에 모든 녹음 음성 데이터의 파형을 정규화하였다.

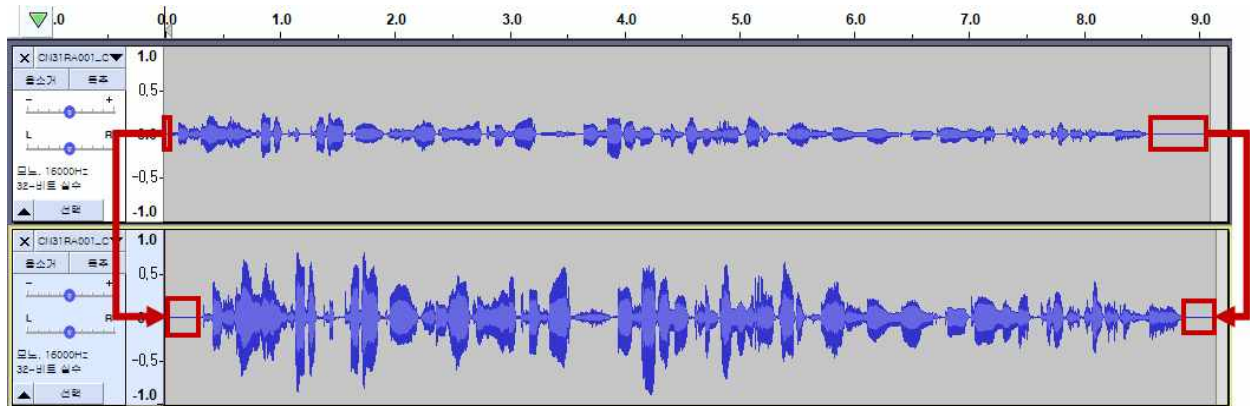


- 녹음 음성 앞뒤 잡음 제거

볼륨 정규화 작업 시 볼륨을 키우거나 줄여 각 음원의 파형을 정규화하게 된다. 볼륨을 줄인 음성들의 경우, 프로그램 실행 후 음원의 파형을 시각적으로 확인하였을 때에는 완전 무음인 상태로 보이더라도 실제 음원 재생 시에는 완전 무음이 아닌 예도 있었다. 이에 음원에서 발화 가장 앞과 뒤에 0.2 ~ 0.5초 기준에 맞는 휴지 구간이 존재하더라도 해당 휴지 구간이 완전한 무음인지는 파악하기 어렵다. 이러한 점에서 원본 녹음 음성이 발화 가장 앞과 뒤에서 기준에 맞는 휴지 구간을 이미 가지고 있더라도, 발화의 가장 앞과 뒤 구간을 모두 잘라내는 작업을 진행하였다.

- 녹음 음성 휴지 구간 삽입

녹음 음성에서 발화 가장 앞과 뒤의 잡음이 제거된 상태에서 0.3초의 완전 무음인 휴지 구간을 일괄적으로 삽입하는 작업을 진행하였다.



- 녹음 음성 시간 정보 정정

녹음 음성에서 발화 가장 앞과 뒤의 잡음 부분을 제거하고, 휴지 구간 0.3초를 삽입함에 따라 변동된 녹음 음성 시간을 정리하여, 메타데이터인 json과 csv의 recordTime 항목을 일괄적으로 정정하는 작업을 진행하였다.

1.10.3 수집·정제 기준

- 녹음 음성 수집 및 검수 기준

- 외국인 녹음 작업자가 녹음한 음성 데이터의 품질을 검수한다.
- 외국인 녹음 작업자가 녹음한 음성의 발음 정확도 및 구문 정확도 등 한국어 발화 정확도는 평가 대상에서 제외된다.
- 아래 기준에 해당하는 경우, 데이터 수집 기준에 맞지 않는 것으로 보아 반려하도록 한다. 반려 후 재작업한 녹음 음성도 같은 이유로 수집 기준에 맞지 않을 경우, 데이터를 제외한다.
- 성실하게 대본을 읽었으나 발음을 알아듣기 힘든 경우와 성실하게 대본을 읽었으나 일부 단어나 문구를 추가하거나 빠뜨리는 등 대본과 다소 차이가 있는 경우는 반려 및 데이터 제거에 해당하지 않는다.

기준	대표 상황
녹음자의 음성이 전혀 녹음되지 않음	TV 소리, 강아지 소리 등 관련 없는 소리만 녹음된 경우
	아무 소리도 녹음되지 않은 경우
녹음자의 음성이 잘 들리지 않음	녹음자의 음성이 너무 작게 들림
	너무 큰 잡음 등으로 인해 전체적으로 녹음자 음성이 들리지 않음
녹음자의 음성 이외에 다른 사람의 음성이 함께 녹음됨	녹음자 음성과 다른 사람의 목소리가 함께 녹음된 경우
	녹음자 음성 뒤로 가사가 있는 노래가 함께 녹음된 경우
	녹음자 음성 뒤로 대사가 있는 TV나 라디오 방송이 함께 녹음된 경우
녹음 음성이 5초 미만	녹음 음성의 녹음 시간이 3초로 기준보다 매우 짧은 경우
녹음자의 음성이 맨 앞이나 맨 뒤에서 잘리게 녹음됨	녹음자의 음성이 맨 앞이나 맨 뒤에서 잘리는 듯이 녹음됨
작업유형에 맞지 않는 작업임	대본 읽기 유형에서 대본에 대한 답변이나 의견만을 녹음한 경우
	질문에 답변하기 유형에서 질문에 대한 답변 없이 질문을 읽는 음성만 녹음한 경우
불성실하게 작업하여 불량 작업임	대본 읽기 유형에서 대본 내용 전체를 아예 다른 내용으로 녹음
	대본 읽기 유형에서 대본 내용 중 한 단어만을 발음하는 등 불성실하게 녹음한 경우
	질문에 답변하기 유형에서 답변 내용이 '네'나 '아니오'와 같이 단답형임
개인정보를 포함	개인을 특정할 수 있는 주민등록번호와 같은 개인정보가 녹음 내용에 포함되어 있음
비속어를 포함	녹음 음성에 차별 및 비하 발언이나 비속어가 녹음된 경우

- 녹음 음성 정제 기준

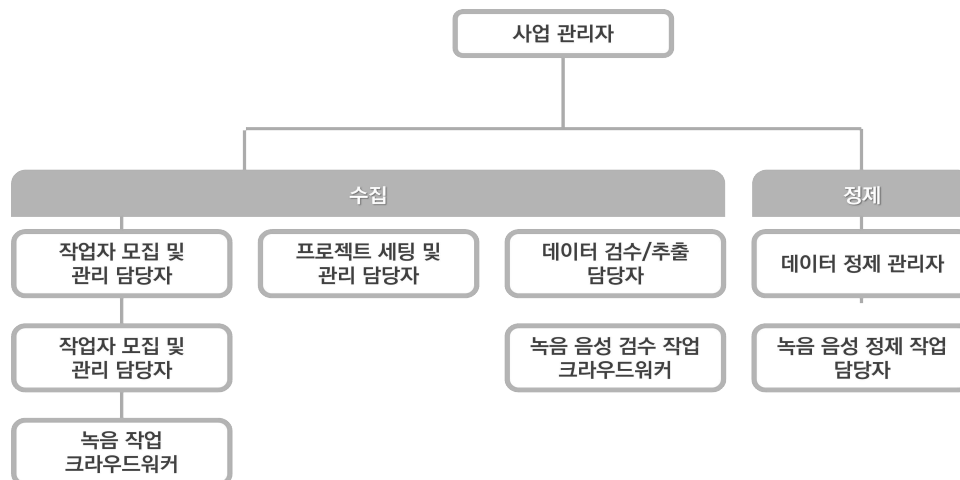
- 음성 데이터의 포맷이 wav인지 확인한다. 샘플레이트가 최소 요구사항인 16k를 충족하는지, 스테레오 채널이 아닌 모노 채널에서 작업이 이루어졌는지 확인한다. 품질 기준과 맞지 않을 경우, 품질 기준에 맞도록 변환한다.
- 녹음 음성 검수 과정에서 미처 걸러지지 않은, 수집 기준에 맞지 않는 음성은 제거한다. 특히, 녹음자의 음성을 알아들을 수 없을 정도의 과도한 소음이 포함된 데이터는 본 과제의 취지와는 맞지 않기 때문에 제거한다.
- 녹음 음성에서 0.2 ~ 0.5초 기준에 충족하는 휴지 구간이 이미 존재하더라도, 완전 무음의 휴지 구간인지 아닌지를 파악하기 어려우므로 발화 음성과 배경음을 구분하여 음성의 맨 앞과 뒤에 있는 잡음을 잘라내 발화 부분만 남기도록 한다. 이후 해당 음성의 맨 앞과 맨 뒤에 각각 0.3초씩 완전 무음의 휴지 구간을 삽입한다.
- 잡음을 자르는 과정에서 잡음과 발화의 크기가 유사하여 뚜렷한 구분이 어려운 경우, 발화 부분이 제거될 위험이 있으므로 무리하게 잡음을 자르지 않는다. 남겨진 잡음 부분 앞에 0.3초의 휴지 구간을 삽입한다.

- 원천데이터 품질 보장 방법

- 음성 데이터의 포맷과 품질에서 AI 알고리즘 학습 및 개발에 보편적으로 이용되는 wav 포맷 16bit 16kHz MONO를 기준으로 하여 전량 수집할 수 있도록 한다.
- 원천 데이터 수집 시 수집 기준에 맞지 않는 작업은 반려하여 녹음 작업자가 기준에 맞추어 다시 작업할 수 있도록 한다. 또한, 주요하게 작업에 오류가 있는 작업자에게는 별도로 주의사항을 안내하도록 하여 오류율을 낮추도록 한다. 이를 통해 정제 과정에 들어가기 전에, 품질이 낮은 데이터의 수를 줄이도록 한다.
- 정제 과정에서 기준에 맞지 않는 음성이 제거될 것을 고려하여, 지속적으로 언어분류와 세트별 음성 수집 시간을 모니터링하여 데이터의 언어분류 및 세트별 목표 시간을 충분히 확보할 수 있도록 한다.

1.10.4 수집·정제 조직

- 수집 및 정제 조직



- 데이터 수집 및 정제를 위한 교육 훈련

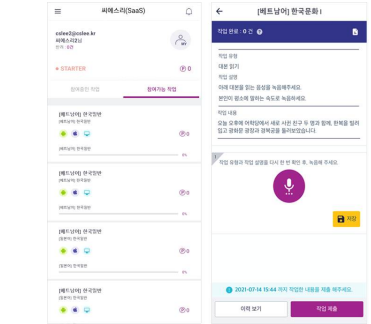
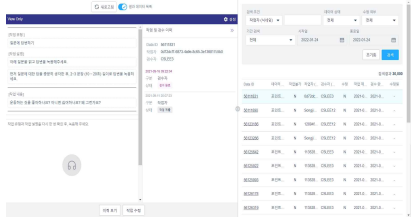

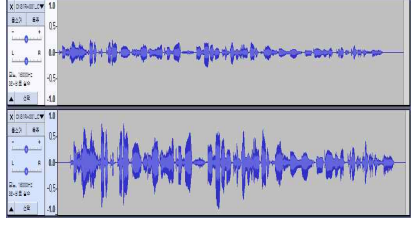
본 사업에서는 원시데이터인 음성을 수집하고, 해당 음성에 대해 외국인의 한국어 발화 특성을 고려하여 전사 작업을 하는 것을 주요 내용으로 한다. 이에 원시데이터인 녹음 음성의 품질에 따라 성과가 달라질 수 있다. 원시데이터 수집 및 정제 시 품질에 영향을 주는 작업은 녹음 작업, 녹음 음성 검수, 녹음 음성 정제로 세 작업이다. 따라서 교육은 녹음 작업자에게 이루어지는 녹음 방법과 녹음 시 품질 기준에 대한 교육과 데이터 검수 및 정제 작업 시 반려 및 제거 기준에 대한 교육 2가지로 진행된다.

구분	내용	방법
녹음 작업 방법	<ul style="list-style-type: none"> 녹음 애플리케이션 사용 및 로그인 방법 녹음 시 지켜야 할 작업 기준 녹음 후 제출 방법 	<ul style="list-style-type: none"> 안내서 배포 유튜브 설명 동영상 제작 및 링크 안내 필요 시 유선상으로 안내 및 문의 해결

녹음 검수 내용	- 녹음 음성 검수 기준 - 녹음 음성 제거 기준	- 가이드라인 배포 및 오프라인 교육 후 작업 결과 검토
녹음 정제 내용	- 녹음 정제 기준 - 녹음 정제 시 수집 기준에 맞지 않는 음성처리기준	- 가이드라인 배포 후 작업 결과 검토

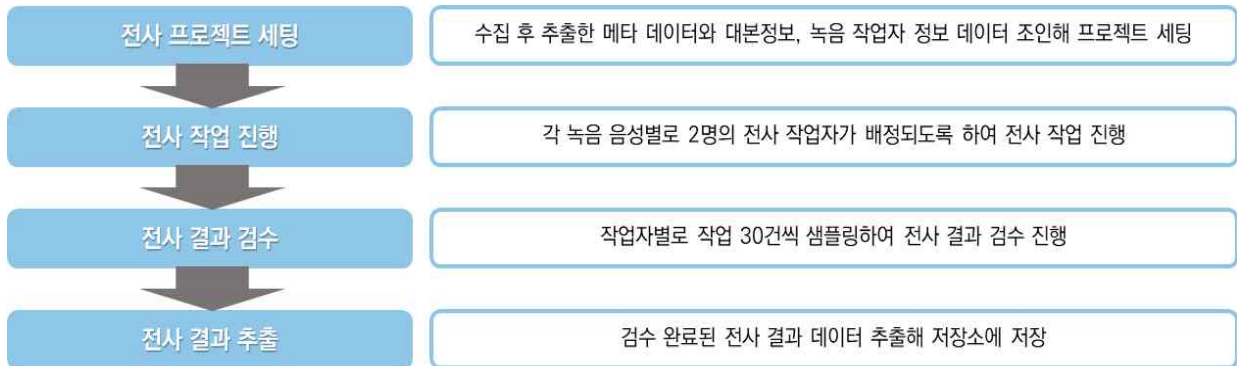
1.10.5 수집·정제 도구

- 데이터 수집 및 정제에는 아래와 같은 도구를 이용하였다. 녹음 음성 수집 및 검수에는 클라우드웍스의 애플리케이션과 웹 페이지 기능을 이용하였다. 또한, 녹음 음성 정제에는 파이썬 pydub 라이브러리를 포함한 다양한 라이브러리를 이용하여 녹음 음성 품질 확인부터 녹음 음성 휴지 구간 삽입까지 자동화하여 진행하였다.

구분	사용 도구	활용 기능	사용 도구 화면 예시
녹음 음성 수집	클라우드웍스 SaaS 어플리케이션	- 언어분류-세트별 녹음 음성 수집 - 녹음 음성 정보 자동 수집 - 녹음 음성 데이터 DB 저장	
녹음 음성 검수	클라우드웍스 SaaS 모니터링 및 검수 화면	- 녹음 음성 수집 현황 모니터링 - 녹음 음성 검수 및 반려 작업	
녹음 정제 내용	파이썬	- 녹음 음성 품질 리샘플링 - 녹음 음성 잡음 구간과 발화 구간 구분 및 잡음 구간 자르기 - 휴지 구간 삽입	
	Audacity	- 녹음 음성 파형 시각적 확인 - 녹음 음성 정제 결과 확인	

1.11 어노테이션/라벨링

1.11.1 어노테이션/라벨링 절차



- 전사 프로젝트 세팅

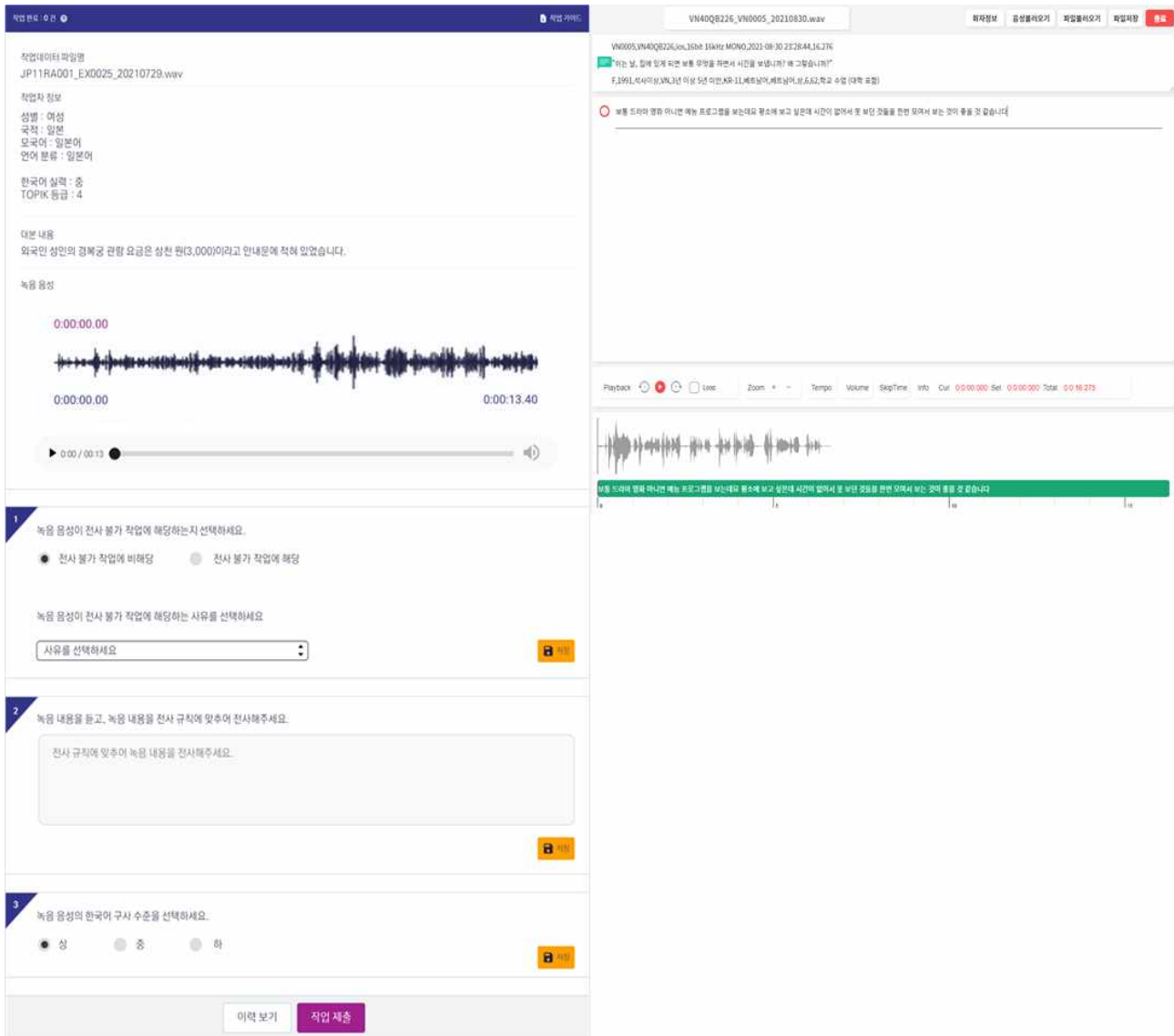
본 과제의 메타데이터 항목은 수집 단계에서 자동으로 입력되거나, 수집한 데이터를 바탕으로 조인하는 과정을 거쳐 구축된다. 녹음 음성 정보 (녹음 단말기 유형, 녹음 품질, 녹음날짜, 녹음 시간)은 데이터 수집 단계에서 자동으로 수집된다. 녹음 작업자 정보에 해당하는 항목(성별, 출생연도, 국적, 국내체류기간 등)은 수집 단계에서 녹음 작업자 등록 시 각 작업자에게 수령한 정보를 바탕으로 구축한 녹음 작업자 정보 데이터를 조인하여 구축한다. 또한, 대본 및 질문 내용도 원시 데이터 선정 단계에서 제작한 대본 및 질문 데이터를 조인하여 구축한다. 이에 따라, 본 과제와 관련해 어노테이션/라벨링 단계에서는 별도로 라벨링 작업을 진행하지 않으며, 전사 작업을 진행한다. 전사란 음성 데이터를 텍스트 데이터로 옮겨 적는 작업을 의미한다.

수집 단계 이후 추출하여 녹음 음성 정보만 기재되어 있는 메타데이터에 대본 및 질문 데이터, 녹음 작업자 정보 데이터를 조인한다. 해당 데이터로 전사 프로젝트를 세팅하도록 한다. 해당 데이터의 정보를 기반으로 전사 작업 후 모든 내용이 기재된 메타데이터를 추출한다.

- 전사 작업 진행

본 사업에서 구축하는 데이터는 주어진 대본 내용을 읽는 대본 읽기 유형의 데이터와 주어진 질문에 대해 자유롭게 답변하는 질문에 답변하기 유형의 데이터로 구분된다. 이에 두 가지 유형에 따라 전사 작업이 진행된다. 대본 읽기 유형은 대본 내용이 주어지며, 전사 작업자는 녹음된 음성을 들으며 대본 내용을 참고하여 전사 작업을 진행한다. 반면 질문에 답변하기 유형의 데이터는 전사 작업자가 녹음된 음성을 듣고 질문 내용을 참고하여 전사 작업을 진행한다. 전사 작업은 클라우드워커들에 의해 수행된다.

대본 읽기 유형 전사 작업자는 전용 웹페이지에 각자 주어진 계정을 통해 접속하여 전사 작업을 진행한다. 질문에 답변하기 유형 전사 작업자는 자체 툴을 이용하여 전사 작업을 진행한다. 두 유형 모두 전사 작업 화면에서 녹음을 위해 녹음 작업자에게 주어진 대본 내용과 녹음 음성, 전사 시 참고할 수 있는 녹음 작업자의 정보 (성별, 국적, 모국어 등)를 확인할 수 있다. 전사 작업자의 주요 작업 내용은 3가지로 녹음 음성 전사 불가 작업에 해당 여부 파악, 녹음 음성 전사, 녹음 내용의 한국어 실력 평가이다. 녹음 음성 전사 불가 작업 해당 여부 판단은 녹음 음성을 듣고 전사 불가 작업 판단 기준에 따라 해당 녹음 음성이 전사 불가 작업에 해당하는지 전사 불가 작업에 해당/비해당 중 하나를 선택하는 작업이다. 전사 불가 작업에 해당할 경우, 관련된 판단 기준을 사유로 리포팅한다. 녹음 음성 전사는 녹음 음성을 듣고 전사 규칙에 따라 전사하는 작업이다. 녹음 음성 한국어 구사 수준 평가는 녹음 음성을 듣고, 한국어 구사 실력 평가 기준에 따라 해당 녹음 음성의 한국어 구사 실력을 상/중/하 중 하나로 선택하는 작업이다. 한국어 구사 실력 평가는 녹음 음성을 단위로 이루어진다.



- 전사 결과 검수

제출된 전사 결과는 모니터링 페이지에서 확인 및 검수 작업을 진행할 수 있다. 작업자별로 작업 내용 중 30건을 샘플링하여, 검수 작업을 진행한다. 검수 과정에서 녹음 음성과 전혀 다른 내용을 전사하였거나, 전사 규칙에 맞지 않는 작업인 경우, 전사 불가 기준에 해당하지 않는 작업이나 전사 불가 작업으로 응답한 경우, 전사 불가 기준에 해당하는 작업이나 전사 불가 작업으로 응답하지 않은 경우 등 전사 가이드라인에 맞지 않는 작업은 해당 전사 작업자가 다시 작업할 수 있도록 반려하였다. 반려 건에 대해 작업자가 재작업을 하면 해당 작업에 대해 재검수를 진행하였다. 단, 검수 시 반려율이 50% 이상으로 높은 작업자는 작업에서 제외하였다. 이와 같은 과정을 통해 전사 가이드라인에 맞지 않는 작업 발생을 줄일 수 있었다.

- 전사 결과 추출

검수 완료된 전사 결과는 정해진 메타데이터 형식에 따라 json과 csv 형식으로 추출하며 전사 단계까지 진행된 데이터임으로 녹음 음성 정보, 전사 결과, 녹음 작업자 정보 모두 기재된 상태로 추출된다.

1.11.2 어노테이션/라벨링 기준

- 전사 불가 작업 해당 여부 판단 기준

- 아래와 같은 기준에 해당하는 녹음 작업은 전사 불가 작업으로 판단한다.
- 전사 불가 작업 해당 여부 판단 기준에 해당하는 녹음 음성의 경우, 본 사업의 목적에 맞지 않는 데이터이므로 제외한다.
- 성실하게 녹음하였지만, 발음을 알아듣기 힘들거나, 성실하게 녹음하였지만 주어진 질문에 대한 답변으로는 적절하지 않은 경우 등 아래 전사 불가 작업 판단 기준에 해당하지 않는 경우는 모두 전사 불가한 작업이 아닌 것으로 판단한다.

전사 불가 작업 기준	전사 불가 작업 해당 여부
녹음자의 음성이 전혀 녹음되지 않아 전사할 수 없는 경우 - TV 소리, 강아지 소리 등 관련 없는 소리만 녹음된 경우 - 아무 소리도 녹음되지 않은 경우	전사 불가 작업에 해당
녹음자의 음성이 잘 들리지 않아 전사 진행에 어려움이 있는 경우 - 녹음자의 음성이 너무 작게 들림 - 너무 큰 잡음 등으로 인해 전체적으로 녹음자 음성이 들리지 않음	전사 불가 작업에 해당
녹음자의 음성 이외에 다른 사람의 음성이 함께 녹음된 경우 - 녹음자 음성과 다른 사람의 목소리가 함께 녹음된 경우 - 녹음자 음성 뒤로 가사가 있는 노래가 함께 녹음된 경우 - 녹음자 음성 뒤로 대사가 있는 TV 방송이나 라디오가 함께 녹음된 경우	전사 불가 작업에 해당
녹음자의 음성이 맨 앞이나 맨 뒤에서 잘리게 녹음되어 정확한 전사가 어려움이 있는 경우 - 녹음자의 음성이 맨 앞이나 맨 뒤에서 음절이 잘리는 듯이 녹음됨	전사 불가 작업에 해당
불성실하게 작업하여, 불량 작업인 경우 - 대본 내용 전체를 아예 다른 내용으로 녹음 - 대본 내용 중 한 단어만을 발음하는 등 명백히 성실하게 녹음한 것으로 보기 어려움 - 질문에 대한 답변을 네/아니오와 같은 단답형으로 한 경우	전사 불가 작업에 해당
답변 내용에 화자 자신 혹은 타인의 개인정보를 포함하여 녹음하였을 경우 - 개인정보 : 주소(시군구만 언급하는 경우는 제외), 주민등록번호, 여권번호, 건강보험번호, 외국인등록번호, 운전면허번호, 자동차번호, 카드번호, 계좌번호(가상계좌 포함), 전화번호, Fax 번호, 이메일 주소, 생년월일 등	전사 불가 작업에 해당

- 녹음 음성 한국어 구사 수준 평가 기준

- 작업자 정보에서 확인할 수 있는 실력이 아닌, 개별 녹음 음성의 한국어 구사 수준을 평가한다.
- 녹음 음성 한국어 구사 수준은 아래와 같은 기준에 따라 상/중/하 중 하나를 선택한다.

구사수준	평가기준
상	- 발음과 억양이 모국어 화자에 준하는 수준으로 명료함 - 개별 음운에는 아직 모국어의 영향이 남아있더라도 모국어 화자에 준하는 수준으로 자연스러움
중	- 어색한 발음과 억양이 나타나지만 말하는 바를 이해하는 데 지장이 없음 - 주저함, 더듬거림, 휴지 등이 때때로 나타나지만 의사소통을 방해하지는 않음
하	- 발음과 억양이 명료하지는 않지만, 주의를 기울이면 말하는 바를 겨우 알아들을 수 있음 - 주저함, 더듬거림, 휴지 등이 자주 나타남

- 전사 기본 원칙

- 외국인의 어눌하고 부정확한 발음을 한국어의 정확한 단어와 매핑하여 인공지능이 외국인의 발음을 인식할 수 있도록 하는 것을 목표로 전사한다.
- 전사 방법은 철자 전사와 발음 전사 2가지로 나뉜다. 철자 전사란 들리는 바를 표준어법에 맞추어 적는 것이며, 발음 전사는 들리는 바를 소리대로 적는 것이다. 본 사업에서는 철자 전사에 따르는 것을 원칙으로 하며, 발음 전사는 진행하지 않는다.
- 전사는 영어, 숫자, 기호, 문장부호 아닌 한글로만 전사하는 것을 원칙으로 한다.
- 대본 읽기 유형은 전사 시 대본 내용을 기본으로 전사 작업을 진행한다. 단, 녹음자가 대본에 없는 내용을 추가해 발화하였거나, 대본 내용 중 빠뜨린 부분이 있다면 전사 규칙에 따라 해당 내용을 반영해 전사한다.

- 대본 읽기 유형 전사 규칙

번호	전사 내용	전사 규칙
1	간투어	<ul style="list-style-type: none"> ○ 간투어란 별다른 의미가 없고 주로 머뭇거림이나 발화 습성으로 인해 나타나는 단어입니다. ○ 간투어는 '이, 그, 저, 어, 아, 예, 음, 응, 엄, 뭐' 로 정의하며, 이외에는 간투어로 전사하지 않습니다. ○ 간투어는 뒤에 '/'를 붙여 전사합니다. ○ 위에서 정의한 간투어에는 포함되지 않으나, 혼잣말이나 웅얼거림 등 의미 없는 내용이 녹음된 경우 해당 부분은 발음한 대로 전사하고, un/을 뒤에 붙여 전사합니다.
2	반복발화	<ul style="list-style-type: none"> ○ 반복발화란 발화 중 더듬거리는 과정에서 단어를 반복해 발화한 경우입니다. ○ 단어의 형태는 바로 뒤의 단어에서 조사를 제외한 부분을 기준으로 판단합니다. 조사를 제외한 바로 뒤 단어와 동일하게 발음을 하였을 경우 완전한 형태로, 일부 음절만 발음하였을 경우 불완전 형태로 판단합니다. ○ 반복발화한 단어가 불완전한 형태일 경우, 반복발화된 부분을 맞춤법에 맞추어 전사하고, 뒤에 '+'를 붙여 정상적인 단어와 구분해 전사합니다. '+' 뒤에는 꼭 공백을 넣어 전사합니다. ○ 반복발화한 단어가 불완전한 형태이고 뒤 단어의 내용과 정확하게 일치하지 않더라도 뒤 단어와 유사한 발음의 단어이며 뒤 단어를 발음하기 위해 더듬거린 내용으로 보이는 경우, '+'를 붙여 전사합니다. ○ 반복발화한 단어가 하나의 단어를 구성할 수 있는 완전한 형태일 경우, 반복발화된 부분을 띄어쓰기를 포함하여 맞춤법에 맞추어 전사하고, 뒤에 '+'는 붙이지 않습니다. ○ 완전한 두 음절 이상이 반복 발화된 경우, '+'를 붙이지 않고 추가 발화처럼 전사합니다. 불완전한 두 음절 이상이 반복발화된 경우, 사전에 등재되지 않은 비표준어이거나 발음을 알아들을 수 없는 부분은 반복 발화규칙이 아니라, 알아듣기 힘든 발화규칙을 참고하여 'un/' 처리하여 전사합니다.
3	추가 발화	<ul style="list-style-type: none"> ○ 추가 발화는 대본에 없는 내용을 녹음자가 추가하여 말한 경우입니다. ○ 추가 발화가 이루어진 경우에는 추가된 내용도 포함하여 맞춤법에 맞게 전사합니다. ○ 사전에 등재되지 않은 비표준어이거나 발음을 알아들을 수 없는 부분이 추가되어 발화된 경우, 알아듣기 힘든 발화규칙을 참고하여 'un/' 처리하여 전사합니다.
4	발화 누락	<ul style="list-style-type: none"> ○ 발화 누락은 대본 내용 중 일부분을 녹음자가 빠뜨리고 녹음한 경우입니다. ○ 발화 누락이 이루어진 경우에는 발화되지 않은 부분을 제외하고 발화된 부분만을 전사 규칙에 맞추어 전사합니다. ○ 단어 중 일부 음절이 누락되어 사전에 등재되지 않은 비표준어로 발화된 경우, 알아듣기 힘든 발화규칙을 참고하여 'un/' 처리하여 전사합니다.
5	대본과 다른 내용 녹음	<ul style="list-style-type: none"> ○ 대본과는 다른 내용으로 대체해 녹음한 경우입니다. ○ 대본과 녹음 내용의 일부분이 다르다면 해당 내용으로 바꾸어 맞춤법에 맞춰 전사하고, 전체가 다르다면 전사 작업은 별도로 하지 않고 전사 작업 불가 대상에 해당하는 것으로 처리하세요. ○ 단, 이 경우는 대체해 녹음한 단어가 우리말 사전에 등재된 표준어이고, 해당 단어를 녹음자가 의도하여 명확하게 발음한 경우이며, 대본 내용을 알아듣기 힘들게 발화한 경우와는 다른 경우입니다.

6	알아듣기 힘든 발화	<ul style="list-style-type: none"> ◦ 녹음 음성에서 명확히 어떤 내용인지 알아듣기 힘든 단어가 있는 경우입니다. ◦ 발음이 불분명하지만, 녹음자가 외국인이라는 점을 고려하였을 때 넓은 의미로 대본 내용을 최대한 유사하게 발음하려 한 것이라고 판단된다면, 대본 내용을 기반으로 맞춤법에 맞추어 전사합니다. ◦ 대본 내용과 다르게 우리말 사전에 등재되지 않은 비표준어를 의도해서 명확하게 발음한 경우, 발음에 최대한 가깝게 전사 후 'un/'을 붙이세요. 'un/'은 어절을 단위로 하여 붙이도록 하며, 'un/'의 앞뒤로는 공백을 꼭 주어 전사합니다. ◦ 대본 내용과 다른 단어를 의도해서 발음하였으며, 무슨 발음을 하는 것인지 알아듣기 힘든 경우 발음에 최대한 가깝게 전사 후 'un/'을 붙이세요. 'un/'은 어절을 단위로 하여 붙이도록 하며, 'un/'의 앞뒤로는 공백을 꼭 주어 전사합니다. ◦ 단, 일상적으로 많이 사용하는 구어체는 비표준어에 해당하지 않는 것으로 보아 un/ 처리하지 않으며 발음 그대로 전사합니다. ◦ 알아듣기 힘들고, 명백히 불성실하게 녹음한 경우에는 별도로 전사 작업을 하지 않고, '전사 작업 불가 대상에 해당'으로 처리하세요.
7	잡음	<ul style="list-style-type: none"> ◦ 녹음 음성에서 녹음자의 음성 이외에 다른 소리가 함께 녹음된 경우입니다. ◦ 화자 잡음은 녹음자가 녹음하는 중간에 내는 잡음입니다. 한숨 소리, 크게 숨을 들이쉬거나 내뿜는 소리, 침 삼키는 소리, 마른 입술소리 등이 있습니다. 외부 잡음은 외부 환경에서 발생한 잡음입니다. 대표적으로는 차 경적 소리와 다른 사람의 말소리가 있습니다. ◦ 외부 잡음은 별도로 전사 처리하지 않으며, 화자 잡음만을 전사 처리합니다. 화자 잡음 중 음성을 듣는 데 방해가 될 정도로 기침 소리, 웃음 소리, 한숨 소리 등 큰 경우에만 전사 처리합니다. ◦ 잡음이 발생한 시점 부분에 'sn/' 을 포함하여 전사하며 'sn/' 앞뒤로는 공백을 주어 전사합니다. 맨 앞에서 잡음이 발생한 경우에는 뒤에만, 맨 뒤에서 발생한 경우에는 앞에만 공백을 주어 전사합니다. ◦ 단, 녹음자의 녹음 내용이 들리지 않을 정도로 과도한 잡음이나, 다른 사람의 말소리가 함께 녹음되었을 경우 해당 작업은 별도로 전사 작업을 하지 않고 '전사 불가 작업에 해당'으로 처리합니다.
8	방언	<ul style="list-style-type: none"> ◦ 표준어가 아닌 방언 발음으로 녹음한 경우입니다. ◦ 방언에 해당하는 발화 내용은 대본 내용의 표준어로 변환하여 전사합니다.
9	외국어/외래어	<ul style="list-style-type: none"> ◦ 외국어 및 외래어는 한국어로 전사하며, 알파벳으로 전사하지 않습니다. ◦ 외국어 및 외래어는 발음이 아닌 한글 표준어법에 맞추어 전사합니다. ◦ 대본에 표기된 괄호 안의 알파벳은 함께 전사하지 않습니다. ◦ 대본 내용을 성실히 녹음하였으나 일부분을 원어 발음으로 발화한 경우, 해당 원어 발음이 국내에서 통상적으로 사용된다면 발음한 그대로를 맞춤법에 맞춰 전사합니다. 해당 원어 발음이 국내에서 통상적으로 사용되지 않는다면 대본 내용에 기반하여 맞춤법에 맞춰 전사합니다.
10	영어 약어	<ul style="list-style-type: none"> ◦ 영어 약어는 발음이 아닌 한글 표준어법에 맞추어 전사합니다. ◦ 한국인 원어민도 혼용하여 사용하는 단어의 경우 발음이 대본 내용과는 다르더라도, 대본 내용에 기반하여 맞춤법에 맞게 전사하세요. ◦ 알파벳 단위로 끊어 읽은 경우, 알파벳 자모 이름 맞춤법 규정에 따라 전사합니다. ◦ 알파벳 자모 이름 맞춤법 규정 A에이/B비/C씨/D디/E이/F에프/G지/H에이치/I아이/J제이/K케이/L엘/ M엠/N엔/O오/P피/Q큐/R알/S에스/T티/U유/V브이또는비/W더블유/X엑스/Y와이/Z지또는제트 ◦ 알파벳 단위로 끊어 읽은 약어는 해당 내용을 모두 띄어쓰기 없이 하나로 전사합니다. ◦ 통상적으로 알파벳 단위로 끊어 읽는 단어를 한 단어로 보고 발화하였을 경우 발음 내용을 표준어로 전사합니다.
11	숫자	<ul style="list-style-type: none"> ◦ 기본적으로 모두 아라비아 숫자 및 기호가 아닌, 한글로 전사합니다. ◦ 문장에 적절하지 않더라도, 한국어 숫자 발음 방법의 하나로 적합할 경우 발화한 내용을 그대로 맞춤법에 맞추어 전사합니다. (예시:1-일-한/2-이-둘/10-십-열-시) ◦ 숫자와 관련된 띄어쓰기는 맞춤법의 [원칙] 규칙을 적용하여 모두 띄어쓰기하여 전사합니다. 이에 따라 단위를 나타내는 '년', '월', '일', '시간', '시', '분', '원' 등은 숫자와 띄어쓰기하여 전사합니다. ◦ 만 단위를 기준으로 띄어쓰기하여 전사합니다. ◦ 중간에 조사가 덧붙여지지 않은 전화번호는 띄어쓰기하지 않고 모두 붙여 전사합니다. ◦ 한국의 특정 기념일을 뜻하는 숫자는 하나로 붙여 전사합니다.

12	단위	<ul style="list-style-type: none"> ○ 맞춤법의 [원칙] 규칙을 적용하여 단위를 나타내는 단어(예 : '리터', '층', '퍼센트' 등)는 숫자와 띄어쓰기 하여 전사합니다. ○ 대본 내용과 다르게 발음하였더라도 통상적으로 읽는 방법에 적합할 경우, 발화한 내용을 그대로 맞춤법에 맞추어 전사합니다. ○ 통상적으로 읽는 방법과 다르게 단위를 읽었을 경우 대본 내용에 기반하여 맞춤법에 맞추어 전사합니다.
13	문장기호	<ul style="list-style-type: none"> ○ 모든 문장기호와 특수문자는 제외하고 전사합니다. (문장부호 유형 : . ? ! ' ' " " ~ 등 특수문자 유형 : + - @ # \$ ^ & 등) ○ 내용상 특수문자의 표기가 필요한 경우, 기호가 아닌 한글로 표기하며 앞말과 띄어 씁니다.
14	띄어쓰기	<ul style="list-style-type: none"> ○ 맞춤법의 [원칙] 규칙을 적용하여 모두 띄어 쓰는 것을 원칙으로 합니다. ○ 맞춤법상으로 볼 때 [원칙] 규칙에 해당하지 않는 띄어쓰기 오류가 대본에 있다면 띄어쓰기 맞춤법 [원칙] 규정에 따라 수정하여 전사합니다.

- 질문에 답변하기 유형 전사 규칙

번호	전사 내용	전사 규칙
1	간투어	<ul style="list-style-type: none"> ○ 간투어란 별다른 의미가 없고 주로 머뭇거림이나 발화 습성으로 인해 나타나는 단어입니다. ○ 간투어는 '이, 그, 저, 어, 아, 에, 음, 응, 엄, 뭐' 로 정의하며, 이외에는 간투어로 전사하지 않습니다. ○ 간투어는 뒤에 '/'를 붙여 전사합니다. ○ 위에서 정의한 간투어에는 포함되지 않으나, 혼잣말이나 웅얼거림 등 의미 없는 내용이 녹음된 경우 해당 부분은 발음한 대로 전사하고, un/을 뒤에 붙여 전사합니다.
2	반복발화	<ul style="list-style-type: none"> ○ 반복발화란 발화 중 더듬거리는 과정에서 단어를 반복해 발화한 경우입니다. ○ 단어의 형태는 바로 뒤의 단어에서 조사를 제외한 부분을 기준으로 판단합니다. 조사를 제외한 바로 뒤 단어와 동일하게 발음을 하였을 경우 완전한 형태로, 일부 음절만 발음하였을 경우 불완전 형태로 판단합니다. ○ 반복발화한 단어가 불완전한 형태일 경우, 반복발화된 부분을 맞춤법에 맞추어 전사하고, 뒤에 '+'를 붙여 정상적인 단어와 구분해 전사합니다. '+' 뒤에는 꼭 공백을 넣어 전사합니다. ○ 반복발화한 단어가 불완전한 형태이고 뒤 단어의 내용과 정확하게 일치하지 않더라도 뒤 단어와 유사한 발음의 단어이며 뒤 단어를 발음하기 위해 떠듬거린 내용으로 보이는 경우, '+'를 붙여 전사합니다. ○ 반복발화한 단어가 하나의 단어를 구성할 수 있는 완전한 형태일 경우, 반복발화된 부분을 띄어쓰기를 포함하여 맞춤법에 맞추어 전사하고, 뒤에 '+'는 붙이지 않습니다. ○ 완전한 두 음절 이상이 반복 발화된 경우, '+'를 붙이지 않고 추가 발화처럼 전사합니다. 불완전한 두 음절 이상이 반복 발화된 경우, 사전에 등재되지 않은 비표준어이거나 발음을 알아들을 수 없는 부분은 반복 발화규칙이 아니라, 알아듣기 힘든 발화규칙을 참고하여 'un/' 처리하여 전사합니다.
3	알아듣기 힘든 발화	<ul style="list-style-type: none"> ○ 녹음 음성에서 명확히 어떤 내용인지 알아듣기 힘든 단어가 있는 경우입니다. ○ 발음이 불분명하고 이상하나 답변을 성실히 녹음한 것으로 보이고, 화자가 외국인이라는 점을 감안할 때 문맥상으로 발음 내용을 파악할 수 있는 경우, 맞춤법에 맞추어 전사하세요. ○ 답변을 성실히 녹음한 것으로 보이지만 발음이 불분명하고 이상하며 화자가 외국인이라는 점을 고려하고 수차례 들어도, 문맥상으로도 판단하기 어려운 경우에는 알아듣기 힘든 부분을 최대한 발음에 가깝게 맞춤법에 맞추어 전사하고, 뒤에 'un/'을 붙이세요. ○ 'un/'은 어절을 단위로 하여 붙이도록 하며, 'un/'의 앞뒤로는 공백을 꼭 주어 전사합니다. ○ 우리말 사전에 등재되지 않은 비표준어를 의도하여 명확하게 발음한 경우, 발음에 최대한 가깝게 전사한 후 'un/'을 붙이세요. ○ 단, 일상적으로 많이 사용하는 구어체는 비표준어에 해당하지 않는 것으로 보아 un/ 처리하지 않으며 발음 그대로 전사합니다.

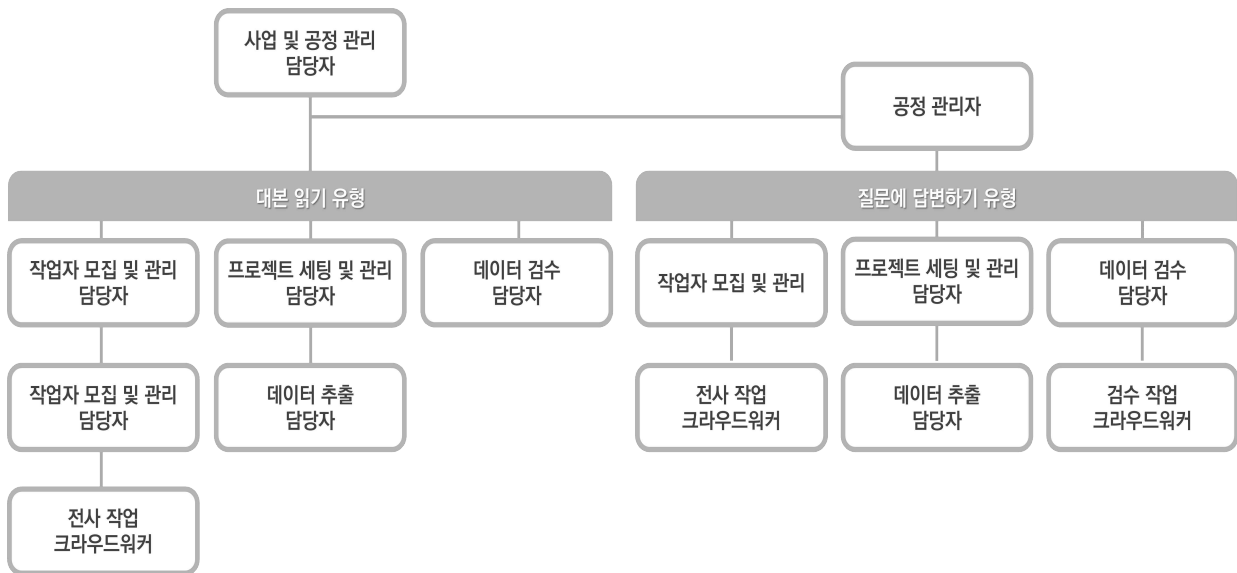
7	잡음	<ul style="list-style-type: none"> ○ 녹음 음성에서 녹음자의 음성 이외에 다른 소리가 함께 녹음된 경우입니다. ○ 화자 잡음은 녹음자가 녹음하는 중간에 내는 잡음입니다. 대표적으로는 웃음소리, 기침 소리 등이 있습니다. ○ 외부 잡음은 외부 환경에서 발생한 잡음입니다. 대표적으로는 차 경적 소리와 다른 사람의 말소리가 있습니다. ○ 외부 잡음은 별도로 전사 처리하지 않으며, 화자 잡음만을 전사 처리합니다. ○ 화자 잡음이 발생하였을 경우, 잡음이 발생한 시점 부분에 'sn/' 을 포함하여 전사합니다. ○ 'sn/' 앞뒤로는 공백을 주어 전사합니다. ○ 단, 녹음자의 녹음 내용이 들리지 않을 정도로 과도한 잡음이나, 다른 사람의 말소리가 함께 녹음되었을 경우 해당 작업은 별도로 전사 작업을 하지 않고 '전사 불가 작업에 해당'으로 처리합니다.
8	방언	<ul style="list-style-type: none"> ○ 표준어가 아닌 방언 발음으로 녹음한 경우입니다. ○ 방언에 해당하는 발화 내용은 표준어로 변환하여 전사합니다.
9	외국어/외래어	<ul style="list-style-type: none"> ○ 외국어 및 외래어는 한국어로 전사하며, 알파벳으로 전사하지 않습니다. ○ 외국어 및 외래어는 발음이 아닌 한글 표준어법에 맞추어 전사합니다. ○ 답변 내용 중 일부분을 원어 발음으로 발화한 경우, 해당 원어 발음이 국내에서 통상적으로 사용된다면 발음한 그대로를 맞춤법에 맞춰 전사합니다. ○ 해당 원어 발음이 국내에서 통상적으로 사용되지 않는다면 표준어법에 맞추어 전사합니다.
10	영어 약어	<ul style="list-style-type: none"> ○ 영어 약어는 발음이 아닌 한글 표준어법에 맞추어 전사합니다. 단, 한국인 원어민도 혼용하여 사용하는 단어의 경우 발음에 가장 가까운 형식으로 전사하세요. ○ 알파벳 단위로 끊어 읽은 경우, 알파벳 자모 이름 맞춤법 규정에 따라 전사합니다. ○ 알파벳 자모 이름 맞춤법 규정 A에이/B비/C씨/D디/E이/F에프/G지/H에이치/I아이/J제이/K케이/L엘/ M엠/N엔/O오/P피/Q큐/R알/S에스/T티/U유/V브이또는비/W더블유/X엑스/Y와이/Z지또는제트 ○ 알파벳 단위로 끊어 읽은 경우 해당 내용을 모두 띄어쓰기 없이 하나로 전사합니다. ○ 통상적으로 알파벳 단위로 끊어 읽는 단어를 한 단어로 보고 발화하였을 경우 발음 내용을 표준어로 전사합니다.
11	숫자	<ul style="list-style-type: none"> ○ 기본적으로 모두 아라비아 숫자 및 기호가 아닌, 한글로 전사합니다. ○ 문장에 적절하지 않더라도, 한국어 숫자 발음 방법의 하나로 적합할 경우 발화한 내용을 그대로 맞춤법에 맞추어 전사합니다. (예시:1-일-한/2-이-둘/10-십-열-시) ○ 숫자와 관련된 띄어쓰기는 맞춤법의 [원칙] 규칙을 적용하여 모두 띄어쓰기하여 전사합니다. 이에 따라 단위를 나타내는 '년', '월', '일', '시간', '시', '분', '원' 등은 숫자와 띄어쓰기하여 전사합니다. ○ 만 단위를 기준으로 띄어쓰기하여 전사합니다. ○ 중간에 조사가 덧붙여지지 않은 전화번호는 띄어쓰기하지 않고 모두 붙여 전사합니다. ○ 한국의 특정 기념일을 뜻하는 숫자는 하나로 붙여 전사합니다.
12	단위	<ul style="list-style-type: none"> ○ 맞춤법의 [원칙] 규칙을 적용하여 단위를 나타내는 단어(예 : '리터', '층', '퍼센트' 등)는 숫자와 띄어쓰기 하여 전사합니다. ○ 통상적으로 읽는 방법에 적합할 경우, 발화한 내용을 그대로 맞춤법에 맞추어 전사합니다. ○ 통상적으로 읽는 방법과 다르게 단위를 읽었을 경우 맞춤법에 맞추어 전사합니다.
13	문장기호	<ul style="list-style-type: none"> ○ 모든 문장기호와 특수문자는 제외하고 전사합니다. (문장부호 유형 : . ? ! ' ' " " ~ 등 특수문자 유형 : + - @ # \$ ^ & 등) ○ 내용상 특수문자의 표기가 필요한 경우에는 기호가 아닌 한글로 표기하며, 앞 말과 띄어씁니다.
14	띄어쓰기	<ul style="list-style-type: none"> ○ 한국어 맞춤법에 기반하여 띄어쓰기하는 것을 원칙으로 합니다. ○ 단, 띄어쓰기 맞춤법에 맞지 않더라도, 발화 내용 중 3초 이상의 틸이 있을 경우 그 부분에 한해 띄어쓰기하여 전사합니다.

- 파일명 부여 규칙

- 파일명은 ① 문장ID, ② 작업자ID, ③ 녹음날짜 순서로 라벨링한다. 파일명은 쌍을 이루는 wav, csv, json 모두 동일하다.
(파일명 예시 : CN21RB002_CN0013_20210803.wav)
- ① 문장ID : 차례대로 언어분류코드, 세트번호, 스크립트번호, 문장번호를 순서대로 나열하여 생성한다.
언어분류 코드는 베트남어 VN, 영어 EN, 일본어 JP, 중국어 CN, 태국어 TH, 기타 EX로 한다.
세트번호는 한국일반 1, 한국생활I 2, 한국생활II 3, 한국문화I 4, 한국문화II 5로 나타낸다.
(문장ID 예시 : CN21RB002)
- ② 작업자ID : 작업자가 해당하는 언어분류 코드와 작업자별 일련번호 4자리를 조합하여 생성한다.
(작업자ID 예시 : CN0013)
- ③ 녹음날짜 : 녹음날짜는 연도, 월, 일 순으로 나타내며, 시간은 포함하지 않는다.

1.11.3 어노테이션/라벨링 조직

- 어노테이션/라벨링 조직



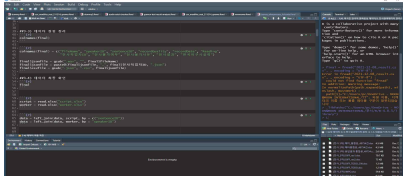
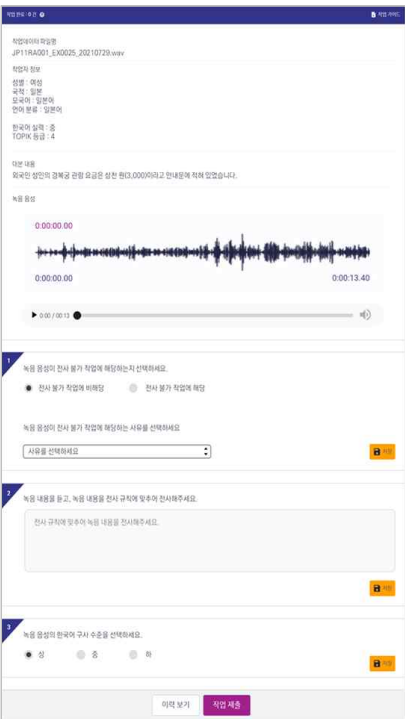

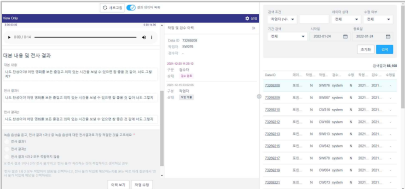
- 데이터 수집 및 정제를 위한 교육 훈련

본 사업의 어노테이션/라벨링 단계에서는 전사 품질을 일관되게 유지하는 것이 중요하다. 이에 전사 작업과 전사 결과 검수 작업과 관련해 교육이 이루어져야 한다. 교육은 전사 작업자에게 이루어진 전사 작업 방법 및 전사 규칙 교육과 전사 결과 검수 작업 시 검수 기준에 대한 교육 2가지로 진행된다.

구분	내용	방법
전사 작업 방법	<ul style="list-style-type: none"> - 전사 웹페이지 사용 및 로그인 방법 - 전사 작업 시 진행할 작업 내용 - 전사 작업 시 지켜야 할 전사 규칙 	<ul style="list-style-type: none"> - 전사 작업 가이드라인 배포 - 온라인 비대면 강의 진행 - 필요 시 이메일 상으로 안내 및 문의 해결
전사 결과 검수 내용	<ul style="list-style-type: none"> - 전사 결과 검수 기준 - 전사 불가 작업 판단 기준 	<ul style="list-style-type: none"> - 가이드라인 배포 후 작업 결과 검토

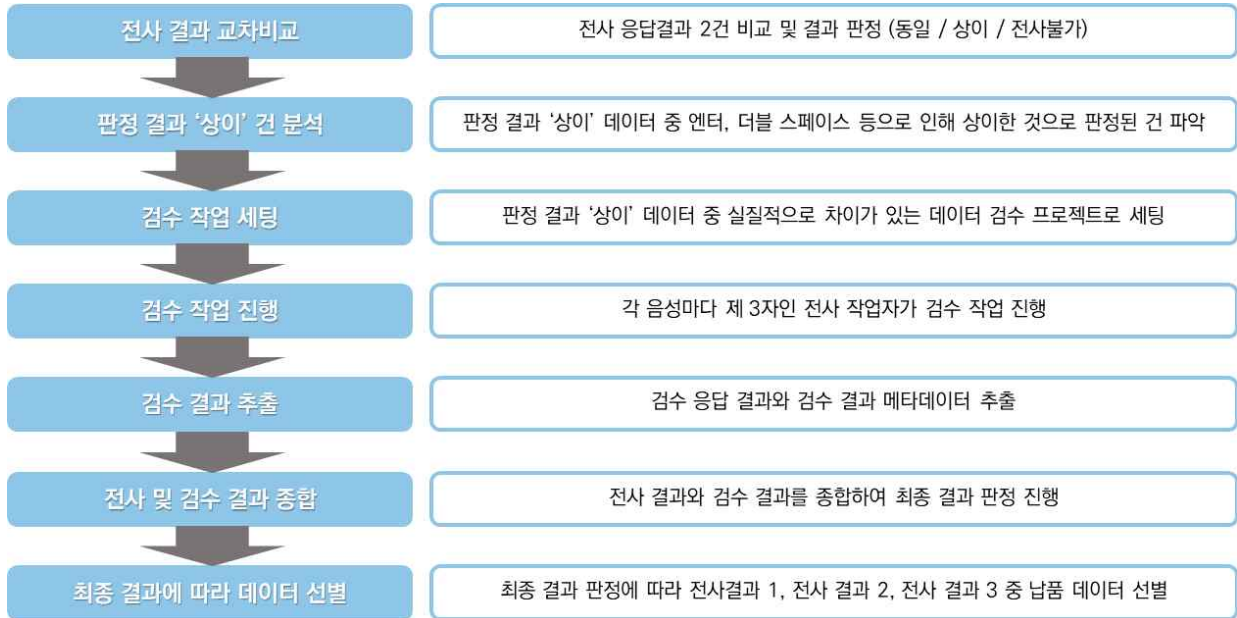
1.11.4 어노테이션/라벨링 도구

- 데이터 어노테이션/라벨링 단계에서는 아래와 같은 도구를 이용하였다. 대본 정보와 녹음 작업자 정보 조인 작업에는 R을 이용하여 전처리를 진행하였다. 대본 읽기 유형의 전사 작업과 전사 결과 검수에는 클라우드웍스 웹페이지와 모니터링 및 검수 화면을 이용하여 진행하였으며, 질문에 답변하기 유형의 전사 작업과 전사 결과 검수는 디그랩 자체 전사 툴을 이용하여 진행하였다.

구분	사용 도구	활용 기능	사용 도구 화면 예시
대본 정보 녹음 작업자 정보 조인	R	<ul style="list-style-type: none"> - 대본 정보와 녹음 작업자 데이터 조인하여 전사 메타데이터 생성의 기본이 되는 데이터 값 기입 - 프로젝트 세팅용 데이터 구성 및 엑셀 데이터로 추출 	
전사 작업	클라우드웍스 SaaS 웹페이지	<ul style="list-style-type: none"> - 전사 작업 진행 - 전사 결과 데이터 DB 저장 	
	디그랩 자체 전사 툴		
전사 결과 검수	클라우드웍스 SaaS 모니터링 및 검수 화면	<ul style="list-style-type: none"> - 전사 작업 현황 모니터링 - 전사 작업 검수 및 반려 작업 	

1.12 검수

1.12.1 검수 절차



- 전사 결과 교차비교

전사 단계 이후 추출하여 녹음 음성 정보, 전사 결과, 녹음 작업자 정보 모두 기재되어 있는 메타데이터를 추출한다. 녹음 음성별로 전사 결과는 2건씩 있으므로, 해당 2건이 동일한지 여부를 1차적으로 비교 판정하도록 한다. 이 때, 두 전사 결과가 한 치의 오차도 없이 동일할 경우 '동일'로, 두 전사 결과에 차이가 있는 경우 '상이'로, 그리고 전사 불가 판단 기준에 해당하여 전사 불가 작업으로 선택된 경우 '전사불가'로 판정한다.

전사결과_1	전사결과_2	1차 판정
고향의 가족한테 특별히 한국의 맛과 아름다움을 알려 드릴 수 있는 선물을 드리고 싶어요	고향의 가족한테 특별히 한국의 맛과 아름다움을 알려 드릴 수 있는 선물을 드리고 싶어요	동일
저는 경복궁에서 삼호선을 타고 교대역에서 이호선으로 갈아타서 집까지 갔습니다	저는 경복궁에서 삼호선을 타고 교대역에서 이호선으로 갈아타서 집까지 갔습니다	동일
우리는 광화문 앞에서 전통 복장을 하고 문을 지키는 아저씨와 함께 사진을 여러 장 찍느라 사진이 가는 줄도 몰랐네요	우리는 광화문 앞에서 전통 복장을 하고 문을 지키는 아저씨와 함께 사진을 여러 장 찍느라 사진이 가는 줄도 몰랐네요	동일
요 만 원 이상 구매 시 삼천 원 할인해 드리고 있으니 잊지 마시고 꼭 사용해 주세요	요 만 원 이상 구매 시 삼천 원 할인해 드리고 있으니 잊지 마시고 꼭 사용해 주세요	동일
가족 모두 여권 보여주시면 감사하겠습니다 고객님의 학생 비자가 있는 거로 나오는데 비자도 보여 주세요	가족 모두 여권 보여주시면 감사하겠습니다 고객님의 학생 비자가 있는 거로 나오는데 비자도 보여 주세요	동일
그럼요 만 원 이상 구매 시 삼천 원 할인해 드리고 있으니 잊지 마시고 꼭 사용해 주세요	그럼요 만 원 이상 구매 시 삼천 원 할인해 드리고 있으니 잊지 마시고 꼭 사용해 주세요	상이
삼오사예요 이공이일에 공이이삼오사입니다 뒷부분도 잘 들으셨어요	삼오사예요 이공이일에 공이이삼오사입니다 뒷부분도 잘 들으셨어요	상이
네 맞습니다 아까 말씀드린 것과 마찬가지로 사 급으로 신청해 주시면 됩니다	네 맞습니다 아까 말씀드린 것과 마찬가지로 사 급으로 신청해 주시면 됩니다	상이
길 안내에 따라 걸어가서 아주 쉽게 한옥같이 생긴 건물 안에 있는 한복 대여소를 찾을 수 있었습니다	길 안내에 따라 걸어가서 아주 쉽게 한옥같이 생긴 건물 안에 있는 한복 대여소를 찾을 수 있었습니다	상이

고향의 가족한테 특별히 한국의 멋과 아름다움을 알려 드릴 수 있는 선물을 드리고 싶어요	고향의 가족한테 특별히 한국의 멋과 아름다움을 알려 드릴 수 있는 선물을 드리고 싶어요	상이
네 감사합니다 여권 여+ 여권 확인 되었습니다 고객님 오늘 어디로 가십니까 혼자 여행하시나요	네 감사합니다 여권 여+ 여권 확인 되었습니다 고객님 오늘 어디로 가십니까 혼자 여행하시나요	상이
네 다음 손님 제가 도와드릴 수 있습니다 안녕하세요 보실 영화는 고르셨나요	네 다음 손님 제가 도와드릴 수 있습니다 안녕하세요 보실 영화 영화는 고르셨나요	상이
미국의 영화 시상식에서 여섯 개 부문 후보에 오르고 배우 윤여정 씨가 여우 조연상을 받은 영화 미나리가 기획에 포함되어 있습니다	미국의 영화 시상식에서 여섯 개 부문 후보에 오르고 배우 윤여정 씨가 유+ 여우 조연상을 받은 영화 미나리가 기획에 포함되어 있습니다	상이
또 취미가 방 청소인 사람도 있어 물건을 버리고 정리하면서 정신이 맑아지고 마음이 가벼워진다는 나는 이해가 안 가	또 취미가 방 청소인 사람도 있어 물건을 버리고 정리하면서 정신이 맑아자+ 맑아지고 마음이 가벼워진다는 나는 이해가 안 가	상이
밤에 불빛을 켜 유등이 강에 떠다니는 것이 얼마나 아름다운 지 몰라요 어두웠던 강이 환하고 반짝반짝 빛났어요	밤에 불빛을 켜 유등이 강에 떠다니는+ 떠다니는 것이 얼마나 나 아름다운지 몰라요 어두웠던 강이 환하고 반짝반짝 빛났어요	상이
고민되네요 아홉 시랑 열 시 중 어/ 조조할인되는 마지막 시간은 언제인가요	고민되네요 아홉 시랑 열 시 중 조조할인되는 마지막 시간은 언제인가요	상이
sn/ 그렇군요 그러면 엽서를 제작해서 거기에 편지를 여러 장 써서 sn/ 우편으로 보내면 되겠어요	그렇군요 그러면 엽서를 제작해서 거기에 편지를 여러 장 써서 우편으로 보내면 되겠어요	상이
여권을 어디에 넣었더라 죄송해요 잠시만요 여기에 제 짐 가방 앞에 꽂혀 있네요 여기 여권 드릴게요	여권을 어디에 넣었더라 죄송해요 잠시만요 여기 제 짐 가방 앞에 꽂혀 있네요 여기 여권 드릴게요	상이
sn/ 제가 제일 좋아하는 색깔은 초록색이어서 안으로 들어가자마+ 들어가자마자 바로 초록색 한복들이 있는 쪽으로 뛰어갔습니다	제가 제일 좋아하는 색깔은 초록색이어서 안으로 들어가자마 자 바로 초록색 한복들이 있는 쪽으로 뛰어갔습니다	상이
우와 아+ 아버님께서 정말 좋아하실 것 같아요 선물은 이엠+ 이엠에스로 보내실 계획이겠지요 가족들이 많이 보고 싶을 것 같아요	우와 아버님께서 정말 좋아하실 것 같아요 선물은 이엠에스로 보내실 계획이겠지요 가족들이 많이 보고 싶을 것 같아요	상이
진이 씨 출장에서 돌아오셨군요! 다시 뵙게 되어 반가워요 그동안 잘 지내셨는지 궁금해요	진이 씨 출장에서 돌아오셨군요 다시 뵙게 되어 반가워요 그동안 잘 지내셨는지 궁금해요	상이
저는 경복궁역에서 삼 호선을 타고 교대역에서 이 호선으로 갈아타서 집까지 갔습니다	저는 경복궁역에서 삼호선을 타고 교대역에서 이호선으로 갈아타서 집까지 갔습니다	상이
다른 방법으로는 두 명을 같이 앉게 해드리고 나머지 한 명만 따로 앉으신다면 좀 더 앞으로 앉게 도와드릴 수 있습니다	다른 방법으로는 두 명 같이 앉게 해드리고 나머지 한 명만 따로 앉으신다면 좀 더 앞으로 앉게 도와드릴 수 있습니다	상이
열 자리로 되어 있는 이 번호를 말씀하시는 거군요 제 번호는 이공이일에 공이이삼오사입니다	열 자리로 되어 있는 이 번호를 말씀하시는 것이군요 제 번호는 이공이일에 공이이삼오사입니다	상이
sn/ 이런 친구들은 자기만의 맛집 지도를 만들고 sn/ 남들에 게 정말 맛있는 식당을 추천하는 걸 sn/ 좋아하더라고	이런 친구들은 자기만의 맛집 지도를 만들고 남들에게 정말 맛있는 식당을 추천하는 걸 좋아하더라고	상이
전사 불가 작업에 해당	전사 불가 작업에 해당	전사불가
전사 불가 작업에 해당	전사 불가 작업에 해당	전사불가
전사 불가 작업에 해당	전사 불가 작업에 해당	전사불가

- 판정 결과 '상이' 건 선별

앞선 단계에서 '상이'로 판정된 데이터 중에는 전사 결과 맨 뒤에서 엔터(/n) 처리하여 문장 내용상으로는 동일한 전사이지만 상이한 것으로 판정된 데이터가 있다. 이에 두 전사 결과의 차이점을 분석하였다. 차이점_1은 전사결과_1이 전사결과_2와 다른 점이며, 차이점_2는 전사결과_2가 전사결과_1과 다른 점이다. 차이점_1이 엔터(/n)이고 차이점_2의 값이 없거나, 차이점_1의 값이 없고 차이점_2가 엔터(/n)이면 두 전사 결과의 차이는 엔터(/n)뿐이다. 이에 해당하는 데이터는 사실상 전사 결과 2건이 동일하다고 볼 수 있다. 이에 2차 판정에서 '검수 X'로 판정하고, 나머지 데이터는 '검수 O'로 판정한다.

전사결과_1	전사결과_2	1차 판정	차이점_1	차이점_2	2차 판정
그럼요 만 원 이상 구매 시 삼천원 할인해 드리고 있으니 잊지 마시고 꼭 사용해 주세요 /n	그럼요 만 원 이상 구매 시 삼천원 할인해 드리고 있으니 잊지 마시고 꼭 사용해 주세요	상이	/n		검수 X
삼오사예요 이공이일에 공이이삼오사입니다 뒷부분도 잘 들으셨어요 /n	삼오사예요 이공이일에 공이이삼오사입니다 뒷부분도 잘 들으셨어요	상이	/n		검수 X
네 맞습니다 아까 말씀드린 것과 마찬가지로 사 급으로 신청해 주시면 됩니다 /n	네 맞습니다 아까 말씀드린 것과 마찬가지로 사 급으로 신청해 주시면 됩니다	상이	/n		검수 X
길 안내에 따라 걸어가서 아주 쉽게 한옥같이 생긴 건물 안에 있는 한복 대여소를 찾을 수 있었습니다	길 안내에 따라 걸어가서 아주 쉽게 한옥같이 생긴 건물 안에 있는 한복 대여소를 찾을 수 있었습니다 /n	상이		/n	검수 X
고향의 가족한테 특별히 한국의 멋과 아름다움을 알려 드릴 수 있는 선물을 드리고 싶어요	고향의 가족한테 특별히 한국의 멋과 아름다움을 알려 드릴 수 있는 선물을 드리고 싶어요 /n	상이		/n	검수 X
네 감사합니다 여+ 여권 확인되었습니다 고객님의 오늘 어디로 가십니까 혼자 여행하시나요	네 감사합니다 여관 여+ 여권 확인되었습니다 고객님의 오늘 어디로 가십니까 혼자 여행하시나요	상이		여관	검수 O
네 다음 손님 제가 도와드릴 수 있습니다 안녕하세요 보실 영화는 고르셨나요	네 다음 손님 제가 도와드릴 수 있습니다 안녕하세요 보실 영화 영화는 고르셨나요	상이		영화	검수 O
미국의 영화 시상식에서 여섯 개 부문 후보에 오르고 배우 윤여정 씨가 여우 조연상을 받은 영화 미나리가 기획에 포함되어 있습니다	미국의 영화 시상식에서 여섯 개 부문 후보에 오르고 배우 윤여정 씨가 유+ 여우 조연상을 받은 영화 미나리가 기획에 포함되어 있습니다	상이		유+	검수 O
또 취미가 방 청소인 사람도 있어 물건을 버리고 정리하면서 정신이 맑아지고 마음이 가벼워진다는 나 는 이해가 안 가	또 취미가 방 청소인 사람도 있어 물건을 버리고 정리하면서 정신이 맑아자+ 맑아지고 마음이 가벼워진다는 나는 이해가 안 가	상이		맑아자+	검수 O
밤에 불빛을 켜 유등이 강에 떠다니는 것이 얼마나 아름다운지 몰라요 어두웠던 강이 환하고 반짝 반짝 빛났어요	밤에 불빛을 켜 유등이 강에 떠다니는+ 떠다니는 것이 얼마나 아름다운지 몰라요 어두웠던 강이 환하고 반짝반짝 빛났어요	상이		떠다니는+	검수 O
고민되네요 아홉 시랑 열 시 중 어/ 조조할인되는 마지막 시간은 언제인가요	고민되네요 아홉 시랑 열 시 중 조조할인되는 마지막 시간은 언제인가요	상이	어/		검수 O
sn/ 그렇군요 그러면 엽서를 제작해서 거기에 편지를 여러 장 써서 sn/ 우편으로 보내면 되겠어요	그렇군요 그러면 엽서를 제작해서 거기에 편지를 여러 장 써서 우편으로 보내면 되겠어요	상이	sn/		검수 O

여권을 어디에 넣었더라 죄송해요 잠시만요 여기에 제 짐 가방 앞에 꽂혀 있네요 여기 여권 드릴게요	여권을 어디에 넣었더라 죄송해 요 잠시만요 여기 제 짐 가방 앞에 꽂혀 있네요 여기 여권 드 릴게요	상이	여기에		검수 O
sn/ 제가 제일 좋아하는 색깔은 초 록색이어서 안으로 들어가자마자 들어가자마자 바로 초록색 한복들 이 있는 쪽으로 뛰어갔습니다	제가 제일 좋아하는 색깔은 초 록색이어서 안으로 들어가자마 자 바로 초록색 한복들이 있는 쪽으로 뛰어갔습니다	상이	sn/		검수 O
우와 아+ 아버님께서 정말 좋아하 실 것 같아요 선물은 이엠+ 이엠 에스로 보내실 계획이겠지요 가족 들이 많이 보고 싶을 것 같아요	우와 아버님께서 정말 좋아하실 것 같아요 선물은 이엠에스로 보내실 계획이겠지요 가족들이 많이 보고 싶을 것 같아요	상이	아+		검수 O
진이 씨 출장에서 돌아오셨군요! 다시 뵙게 되어 반가워요 그동안 잘 지내셨는지 궁금해요	진이 씨 출장에서 돌아오셨군요! 다시 뵙게 되어 반가워요 그동 안 잘 지내셨는지 궁금해요	상이	돌아오셨군요!	돌아오셨군요	검수 O
저는 경북공역에서 삼 호선을 타 고 교대역에서 이 호선으로 갈아 타서 집까지 갔습니다	저는 경북공역에서 삼호선을 타 고 교대역에서 이호선으로 갈아 타서 집까지 갔습니다	상이	삼	삼호선을	검수 O
다른 방법으로는 두 명을 같이 앉 게 해드리고 나머지 한 명만 따로 앉으신다면 좀 더 앞으로 앉게 도 와드릴 수 있습니다	다른 방법으로는 두 명 같이 앉 게 해드리고 나머지 한 명만 따 로 앉으신다면 좀 더 앞으로 앉 게 도와드릴 수 있습니다	상이	명을	명	검수 O
열 자리로 되어 있는 이 번호를 말씀하시는 거군요 제 번호는 이 공이일에 공이이삼오사입니다	열 자리로 되어 있는 이 번호를 말씀하시는 것이군요 제 번호는 이공이일에 공이이삼오사입니다	상이	거군요	것이군요	검수 O
sn/ 이런 친구들은 자기만의 맛집 지도를 만들고 sn/ 남들에게 정말 맛있는 식당을 추천하는 걸 sn/ 좋아하더라고	이런 친구들은 자기만의 맛집 지도를 만들고 남들에게 정말 맛있는 식당을 추천하는 걸 좋 아하더라고	상이	sn/	식당을	검수 O

- 검수 작업 세팅

앞서 2차 판정 시 '검수 O'로 판정된 데이터만을 선별하여 검수 작업에 세팅한다. 녹음 음성별로 전사 결과 2건을 조인하여, 검수 작업 시 전사 결과 2건 모두를 확인할 수 있도록 세팅 데이터를 구성한다. 검수 작업은 녹음 음성별로 검수 작업자 1명이 배정되어 작업을 진행할 수 있도록 한다.

- 검수 작업 진행

본 사업에서 구축하는 데이터 유형이 대본 읽기와 질문에 답변하기 두 가지로 나누어져 있으므로, 검수 작업도 유형에 따라 진행된다. 대본 읽기 유형의 검수 작업 주요 내용은 4가지로 전사 결과 중 적절한 결과 선택, 녹음 음성 전사 불가 작업 해당 여부 파악, 녹음 음성 전사, 녹음 내용의 한국어 실력 평가이다. 검수 작업자는 대본 내용과 전사 결과 2건을 확인하고, 전사 결과 중 녹음 음성의 전사 결과로 더욱 적절한 전사 결과를 고르는 작업을 진행한다. 녹음 음성이 앞선 구축 단계에서 미처 걸러지지 못한 전사 불가 건일 경우, 전사 불가 작업에 해당하는 것으로 선택한다. 또한, 두 전사 결과가 모두 적절하지 않을 경우 주어진 대본 내용과 이전 전사 결과를 참고하여 녹음 음성을 듣고 새로운 전사 결과를 작성하도록 한다. 더불어, 새로운 전사 결과 작성 시 한국어 구사 실력 또한 새롭게 상/중/하 중 하나를 선택하도록 한다.

질문에 답변하기 유형의 검수 작업 주요 내용은 4가지로 전사 결과가 적절한지 여부 선택, 전사 결과 부적절 사유 선택, 전사 불가 작업 판단 기준 선택, 녹음 음성 전사, 녹음 내용의 한국어 실력 평가이다. 검수 작업자는 질문 내용과 전사 결과 1건을 확인하고, 전사 결과가 녹음 음성을 전사한 결과로 적절한지 여부를 고르는 작업을 진행한다. 녹음 음성이 앞선 구축 단계에서 미처 걸러지지 못한 전사 불가 건이라면 전사 불가로 선택하여 응답하도록 한다. 또한, 제시된 전사 결과가 적절하지 않을 경우 주어진 질문 내용과 이전 전사 결과를 참고하여 녹음 음성을 듣고 새로운 전사 결과를 작성하도록 한다. 새로운 전사 결과 작성 시 한국어 구사 실력 또한 새롭게 상/중/하 중 하나를 선택하도록 한다.

작업 완료 : 0 분	작업 기록	작업 완료 : 0 분	작업 기록
<p>작업데이터 확인명 JP11RA001_JP0025_20210729.wav</p> <p>작업자 정보 성별 : 여성 국적 : 일본 모국어 : 일본어 언어 분류 : 일본어 한국어 실력 : 중 TOPIK 등급 : 4</p> <p>녹음 음성</p> <p>0:00:00.00 0:00:00.00 0:00:13.40</p> <p>▶ 0:00 / 00:13</p>		<p>작업데이터 확인명 JP11QB001_JP0025_20210729.wav</p> <p>작업자 정보 성별 : 여성 국적 : 일본 모국어 : 일본어 언어 분류 : 일본어 한국어 실력 : 중 TOPIK 등급 : 4</p> <p>녹음 음성</p> <p>0:00:00.00 0:00:00.00 0:00:13.40</p> <p>▶ 0:00 / 00:13</p>	
<p>대본 내용 및 전사 결과</p> <p>대본 내용 외국인 성인의 경제공 관할 요금은 삼천 원(3,000)이라고 안내문에 적혀 있습니다.</p> <p>전사 결과 1 외국인 성인의 경제공 관할 요금은 삼천 원이라고 안내문에 적혀 있었* 있었습니다.</p> <p>전사 결과 2 외국인 성인의 경제공 관할 요금은 삼천 원이라고 함/함* 안내문에 적혀 있었* 있었습니다.</p>		<p>대본 내용 및 전사 결과</p> <p>대본 내용 불 먹거나 한 먹는 음식이 있습니까? 그 이유는 무엇입니까?</p> <p>전사 결과 저는 순댓국이라 그 원장 한 먹어요 냄새가 너무 심해요</p>	
<p>1 녹음 음성을 듣고, 전사 결과 1과 2 중 녹음 음성에 대한 전사 결과로 가장 적절한 것을 고르세요. ※ 전사 결과 1이나 2가 "전사 불가"이고 "전사 불가" 처리하는 것이 적합하다고 생각하실 경우, "전사 결과 1과 2 모두 적합하지 않음"을 선택하세요.</p> <p><input checked="" type="radio"/> 전사 결과 1 <input type="radio"/> 전사 결과 2 <input type="radio"/> 전사 결과 1과 2 모두 적합하지 않음</p>		<p>1 녹음 음성을 듣고, 전사 결과가 녹음 음성에 대한 전사 결과로 적절한지 선택하세요.</p> <p><input checked="" type="radio"/> 적절하다 <input type="radio"/> 부적절하다 ※ 적절한 전사 결과는 녹음 내용과 일치하며, 전사 불가 처리 기준에 해당하지 않고, 주어진 전사 규칙에 적합한 결과입니다.</p> <p>전사 결과가 녹음 음성에 대한 전사 결과로 부적절하다고 생각하시는 사유를 선택하세요</p> <p>사유를 선택하세요</p>	
<p>2 녹음 음성이 전사 불가 작업에 해당하는지 선택하세요.</p> <p><input checked="" type="radio"/> 전사 불가 작업에 해당함 <input type="radio"/> 전사 불가 작업에 해당</p> <p>녹음 음성이 전사 불가 작업에 해당하는 사유를 선택하세요</p> <p>사유를 선택하세요</p>		<p>2 사유로 "전사 불가 작업에 해당"을 선택하셨습니다. 녹음 음성이 전사 불가 작업에 해당하는 사유를 선택하세요.</p> <p>사유를 선택하세요</p>	
<p>3 녹음 내용을 듣고, 녹음 내용을 전사 규칙에 맞추어 전사하세요.</p> <p>녹음 음성 (위 녹음 음성과 동일)</p> <p>▶ 0:00 / 00:13</p> <p>전사 규칙에 맞추어 녹음 내용을 전사하세요</p>		<p>3 녹음 내용을 듣고, 녹음 내용을 전사 규칙에 맞추어 전사하세요.</p> <p>녹음 음성 (위 녹음 음성과 동일)</p> <p>0:00:00.00 0:00:00.00 0:00:13.40</p> <p>▶ 0:00 / 00:13</p> <p>전사 규칙에 맞추어 녹음 내용을 전사하세요</p>	
<p>녹음 음성 한국어 구사 수준 선택 결과</p> <p>녹음 음성 한국어 구사 수준 선택 결과 1 상</p> <p>녹음 음성 한국어 구사 수준 선택 결과 2 중</p>		<p>녹음 음성 한국어 구사 수준 선택 결과</p> <p>녹음 음성 한국어 구사 수준 선택 결과 1 상</p>	
<p>4 녹음 음성 한국어 구사 수준을 선택하세요</p> <p><input checked="" type="radio"/> 상 <input type="radio"/> 중 <input type="radio"/> 하</p>		<p>4 녹음 음성 한국어 구사 수준을 선택하세요</p> <p><input checked="" type="radio"/> 상 <input type="radio"/> 중 <input type="radio"/> 하</p>	
<p>이력 보기 작업 재출</p>		<p>이력 보기 작업 재출</p>	

- 검수 결과 추출

제출 완료된 검수 작업 결과는 검수 작업 응답 결과.xlsx 데이터와 메타데이터 형식에 따른 json, csv 데이터로 추출한다. 정해진 메타데이터 형식에 따라 추출하는 json과 csv 데이터는 녹음 음성 정보, 전사 결과, 녹음 작업자 정보 모두가 기재된 상태로 추출하도록 한다.

- 전사 및 검수 결과 종합

검수 작업 응답 결과를 토대로 작업 결과 중 최종 납품할 데이터를 판정하도록 한다. 검수 작업에서 전사 결과 1이 적절하다는 응답이었을 경우, 전사 결과 1을 선택하며 전사 결과 2가 적절하다는 응답이었을 경우 전사 결과 2를 선택하는 것으로 판정한다. 새롭게 전사한 내용이 있을 경우에는 새롭게 전사한 전사 결과 3을 선택하는 것으로 판정한다. 또한, 검수 작업 중 전사 불가 작업에 해당하는 것으로 응답한 데이터는 전사 불가로 판정한다.

전사결과_1	전사결과_2	1차 판정	2차 판정	적절한 전사	전사결과_3	최종납품 결과 선택
네 감사합니다 여+ 여권 확인되었습니다 고객님의 오늘 어디로 가십니까 혼자 여행하시나요	네 감사합니다 여관 여+ 여권 확인되었습니다 고객님의 오늘 어디로 가십니까 혼자 여행하시나요	상이	검수 O	전사결과_1		전사결과_1
네 다음 손님 제가 도와드릴 수 있습니다 안녕하세요 보실 영화는 고르셨나요	네 다음 손님 제가 도와드릴 수 있습니다 안녕하세요 보실 영화 영화는 고르셨나요	상이	검수 O	전사결과_2		전사결과_2
미국의 영화 시상식에서 여섯 개 부문 후보에 오르고 배우 윤여정 씨가 여우 조연상을 받은 영화 미나리가 기획에 포함되어 있습니다	미국의 영화 시상식에서 여섯 개 부문 후보에 오르고 배우 윤여정 씨가 유+ 여우 조연상을 받은 영화 미나리가 기획에 포함되어 있습니다	상이	검수 O	전사결과_2		전사결과_2
또 취미가 방 청소인 사람도 있어 물건을 버리고 정리하면서 정신이 맑아지고 마음이 가벼워진다나 나는 이해가 안가	또 취미가 방 청소인 사람도 있어 물건을 버리고 정리하면서 정신이 맑아자+ 맑아지고 마음이 가벼워진다나 나는 이해가 안가	상이	검수 O	없음	또 취미가 방 청소인 사람도 있어 물건을 버리고 정리하면서 정신이 맑아지+ 맑아지고 마음이 가벼워진다나 나는 이해가 안가	전사결과_3
밤에 불빛을 켜 유등이 강에 떠다니는 것이 얼마나 아름다운지 몰라요 어두웠던 강이 환하고 반짝반짝 빛났어요	밤에 불빛을 켜 유등이 강에 떠다니는+ 떠다니는 것이 얼마나 아름다운지 몰라요 어두웠던 강이 환하고 반짝반짝 빛났어요	상이	검수 O	전사결과_2		전사결과_2
고민되네요 아홉 시랑 열 시 중 여/ 조조할인되는 마지막 시간은 언제인가요	고민되네요 아홉 시랑 열 시 중 조조할인되는 마지막 시간은 언제인가요	상이	검수 O	전사결과_1		전사결과_1
sn/ 그렇군요 그러면 엽서를 제작해서 거기에 편지를 여러 장 써서 sn/ 우편으로 보내면 되겠어요	그렇군요 그러면 엽서를 제작해서 거기에 편지를 여러 장 써서 우편으로 보내면 되겠어요	상이	검수 O	없음	그렇군요 그러면 엽서를 제작해서 거기에 편지를 여러 장 써서 우+ 우편으로 보내면 되겠어요	전사결과_3
여권을 어디에 넣었더라 죄송해요 잠시만요 여기에 제 짐 가방 앞에 꽂혀 있네요 여기 여권 드릴게요	여권을 어디에 넣었더라 죄송해요 잠시만요 여기에 제 짐 가방 앞에 꽂혀 있네요 여기 여권 드릴게요	상이	검수 O	전사결과_1		전사결과_1

sn/ 제가 제일 좋아하는 색깔은 초록색이어서 안으로 들어가자마자 들어가지마자 바로 초록색 한복들이 있는 쪽으로 뛰어갔습니다	제가 제일 좋아하는 색깔은 초록색이어서 안으로 들어가자마자 바로 초록색 한복들이 있는 쪽으로 뛰어갔습니다	상이	검수 O	전사불가		전사불가
우와 아+ 아버님께서 정말 좋아하실 것 같아요 선물은 이엠+ 이엠에스로 보내실 계획이겠지요 가족들이 많이 보고 싶을 것 같아요	우와 아버님께서 정말 좋아하실 것 같아요 선물은 이엠에스로 보내실 계획이겠지요 가족들이 많이 보고 싶을 것 같아요	상이	검수 O	전사결과_1		전사결과_1
진이 씨 출장에서 돌아오셨군요! 다시 뵙게 되어 반가워요 그동안 잘 지내셨는지 궁금해요	진이 씨 출장에서 돌아오셨군요! 다시 뵙게 되어 반가워요 그동안 잘 지내셨는지 궁금해요	상이	검수 O	전사결과_2		전사결과_2
저는 경복궁역에서 삼호선을 타고 교대역에서 이호선으로 갈아타서 집까지 갔습니다	저는 경복궁역에서 삼호선을 타고 교대역에서 이호선으로 갈아타서 집까지 갔습니다	상이	검수 O	없음	저는 경복궁역에서 삼호선을 타고 교대역에서 이호선으로 갈아타서 sn/ 집까지 갔습니다	전사결과_3
다른 방법으로는 두 명을 같이 앉게 해드리고 나머지 한 명만 따로 앉으신다면 좀 더 앞으로 앉게 도와드릴 수 있습니다	다른 방법으로는 두 명이 같이 앉게 해드리고 나머지 한 명만 따로 앉으신다면 좀 더 앞으로 앉게 도와드릴 수 있습니다	상이	검수 O	전사결과_1		전사결과_1
열 자리로 되어 있는 이 번호를 말씀하시는 거군요 제 번호는 이공이일에 공이이삼오사입니다	열 자리로 되어 있는 이 번호를 말씀하시는 것이군요 제 번호는 이공이일에 공이이삼오사입니다	상이	검수 O	없음	열 자리로 되어 있는 이 번호를 말+ 말씀하시는 것이군요 제 번호는 이공이일에 공이이삼오사입니다	전사결과_3
sn/ 이런 친구들은 자기만의 맛집 지도를 만들고 sn/ 남들에게 정말 맛있는 식당을 추천하는 걸 sn/ 좋아하더라고	이런 친구들은 자기만의 맛집 지도를 만들고 남들에게 정말 맛있는 식당을 추천하는 걸 좋아하더라고	상이	검수 O	전사불가		전사불가

또한, 앞서 판정결과 1에서 '동일'로 판정된 건과, '상이' 건 중 '검수X'로 판정된 건도 마찬가지로 최종 납품할 데이터를 판정하도록 한다. '동일'로 판정된 건은 전사 결과 1을 선택하는 것으로 판정하며, '상이' 건 중 '검수 X'로 판정된 건은 전사 결과 1과 전사 결과 2 중 엔터 처리가 되지 않은 데이터를 선택하는 것으로 판정한다.

전사결과_1	전사결과_2	1차 판정	차이점_1	차이점_2	2차 판정	최종납품결과 선택
고향의 가족한테 특별히 한국의 멋과 아름다움을 알려 드릴 수 있는 선물을 드리고 싶어요	고향의 가족한테 특별히 한국의 멋과 아름다움을 알려 드릴 수 있는 선물을 드리고 싶어요	동일			-	전사결과_1
저는 경복궁역에서 삼호선을 타고 교대역에서 이호선으로 갈아타서 집까지 갔습니다	저는 경복궁역에서 삼호선을 타고 교대역에서 이호선으로 갈아타서 집까지 갔습니다	동일			-	전사결과_1

우리는 광화문 앞에서 전통 복장을 하고 문을 지키는 아저씨와 함께 사진을 여러 장 찍느라 사진이 가는 줄도 몰랐네요	우리는 광화문 앞에서 전통 복장을 하고 문을 지키는 아저씨와 함께 사진을 여러 장 찍느라 사진이 가는 줄도 몰랐네요	동일			-	전사결과_1
요 만 원 이상 구매 시 삼천원 할인해 드리고 있으니 잊지 마시고 꼭 사용해 주세요	요 만 원 이상 구매 시 삼천원 할인해 드리고 있으니 잊지 마시고 꼭 사용해 주세요	동일			-	전사결과_1
가족 모두 여권 보여주시면 감사하겠습니다 고객님의 학생비자가 있는 거로 나오는데 비자도 보여 주세요	가족 모두 여권 보여주시면 감사하겠습니다 고객님의 학생비자가 있는 거로 나오는데 비자도 보여 주세요	동일			-	전사결과_1
그럼요 만 원 이상 구매 시 삼천 원 할인해 드리고 있으니 잊지 마시고 꼭 사용해 주세요 /n	그럼요 만 원 이상 구매 시 삼천 원 할인해 드리고 있으니 잊지 마시고 꼭 사용해 주세요 /n	상이	/n		검수 X	전사결과_2
삼오사예요 이공이일에 공이이 삼오사입니다 뒷부분도 잘 들으셨어요 /n	삼오사예요 이공이일에 공이이 삼오사입니다 뒷부분도 잘 들으셨어요	상이	/n		검수 X	전사결과_2
네 맞습니다 아까 말씀드린 것과 마찬가지로 사 급으로 신청해 주시면 됩니다 /n	네 맞습니다 아까 말씀드린 것과 마찬가지로 사 급으로 신청해 주시면 됩니다	상이	/n		검수 X	전사결과_2
길 안내에 따라 걸어가서 아주 쉽게 한옥같이 생긴 건물 안에 있는 한복 대여소를 찾을 수 있었습니다	길 안내에 따라 걸어가서 아주 쉽게 한옥같이 생긴 건물 안에 있는 한복 대여소를 찾을 수 있었습니다 /n	상이		/n	검수 X	전사결과_1
고향의 가족한테 특별히 한국의 멋과 아름다움을 알려 드릴 수 있는 선물을 드리고 싶어요	고향의 가족한테 특별히 한국의 멋과 아름다움을 알려 드릴 수 있는 선물을 드리고 싶어요 /n	상이		/n	검수 X	전사결과_1
전사 불가 작업에 해당	전사 불가 작업에 해당	전사불가			-	전사불가
전사 불가 작업에 해당	전사 불가 작업에 해당	전사불가			-	전사불가
전사 불가 작업에 해당	전사 불가 작업에 해당	전사불가			-	전사불가

- 최종 결과에 따라 데이터 선별

최종 납품할 데이터로 판정한 결과에 따라 각각 전사 결과 1, 전사 결과 2, 전사 결과 3 중 선별해 지정된 폴더로 복사 및 이동하도록 한다. 해당 데이터는 최종 납품할 데이터로, 기존 전사와 검수 결과 데이터와는 별도로 관리하도록 한다.

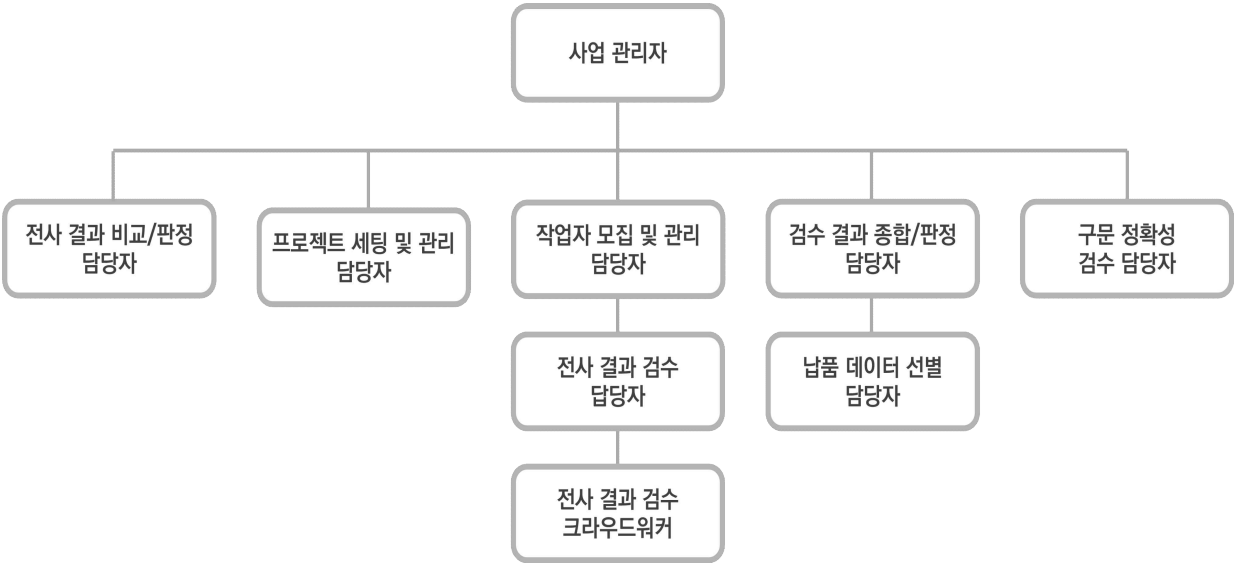
1.12.2 검수 기준

- 검수 기준

- 검수 작업은 녹음 음성을 전사한 결과가 적절한지를 중점적으로 확인하는 것으로 한다.
- 적절한 전사 결과는 녹음 음성을 전사한 결과이며, 전사 규칙에 따라 올바르게 전사된 내용이어야 한다.
- 적절한 전사 결과에는 줄바꿈(엔터 처리)와 띄어쓰기가 두 번 이루어진 더블 스페이스가 있어서는 안 된다.
- 검수 시 전사 결과가 적절한지 판단하는 기준은 본 사업의 전사 규칙이다. 또한, 전사 규칙은 전사 단계에서 정의한 대본 읽기 유형 전사 규칙과 질문에 답변하기 전사 규칙과 동일하게 적용한다.
- 검수 시 녹음 음성이 전사 불가 작업에 해당하는지를 판단하는 기준은 전사 단계에서 정의한 전사 불가 작업 판단 기준과 동일한 내용으로 한다.

1.12.3 검수 조직

- 검수 조직



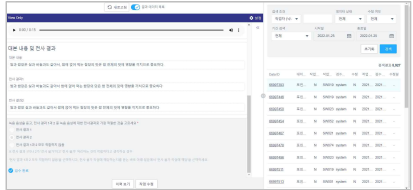
- 데이터 수집 및 정제를 위한 교육 훈련

본 사업의 검수 단계에서는 검수 품질을 일관되게 유지하는 것이 중요하다. 이에 검수 작업과 관련해 교육이 이루어져야 한다. 교육은 검수 작업자에게 이루어진 검수 작업 방법 및 전사 규칙 교육이 진행되었다.

구분	내용	방법
검수 작업 방법	<ul style="list-style-type: none">- 검수 웹페이지 사용 및 로그인 방법- 검수 작업 시 진행할 작업 내용- 검수 작업 시 지켜야 할 전사 규칙	<ul style="list-style-type: none">- 검수 작업 가이드라인 배포- 필요 시 이메일 상으로 안내 및 문의 해결

1.12.4 검수 도구

구분	사용 도구	활용 기능	사용 도구 화면 예시
전사 결과 비교/판정 및 검수 데이터 생성	R	<ul style="list-style-type: none">- 전사 결과 1과 전사 결과 2 비교 및 판정 진행- 검수 작업 세팅용 데이터 생성	
검수 작업	클라우드웍스 SaaS 웹페이지	<ul style="list-style-type: none">- 검수 작업 진행- 검수 결과 데이터 DB 저장	

검수 결과 모니터링	클라우드웍스 SaaS 모니터링 화면	- 검수 작업 현황 모니터링	
---------------	------------------------	-----------------	---

1.12.5 기타 품질관리 활동

- 컨소시엄 내 긴밀한 피드백
 - 데이터 구축 중 발생하는 품질 관련 이슈를 컨소시엄 주간 회의 시 전달하여 컨소시엄 내 모든 기관이 해당 이슈에 대해 인지하도록 하고, 피드백을 통해 품질 관련 이슈를 방지 및 해결할 수 있도록 한다.
 - 모든 공정이 동시에 진행되는 만큼, 앞의 공정에서 이슈가 생길 경우 유기적으로 공정 진행 일정이 늦어질 수 있으므로 컨소시엄과의 긴밀한 소통을 통해 공정 진행 일정을 조율한다.
- 클라우드워커 오픈 채팅방 운영
 - 클라우드워커의 작업이 진행되는 수집, 전사, 검수 공정에서는 작업자 오픈채팅방을 운영하여 작업 방법과 전사/검수 규칙에 대한 작업자들의 문의를 실시간으로 해결하여 작업 시 정확한 규칙을 토대로 작업할 수 있도록 한다. 또한, 작업 중 기술적 오류나 전사 규칙과 관련해 중요한 공지사항이 있을 경우 오픈채팅방을 통해 공지하여 작업자들이 빠르게 공지 내용을 확인하고 작업에 적용할 수 있도록 한다.
- 자체 평가 및 평가 결과 반영
 - 검수 단계 이후 최종 납품 데이터에 대해 자체적으로 정확성 검사를 진행하여 음성 재생 불가, csv와 json 쌍의 데이터 불일치, csv와 json 데이터 내 데이터값 누락 및 추가, 데이터값의 유효성 등에 대해 검토한다. 또한, 이와 관련하여 이슈가 있을 경우, 해당 문제에 대해 정정 작업을 거쳐 보다 정확성을 높일 수 있도록 한다.

1.13 활용

1.13.1 활용 모델

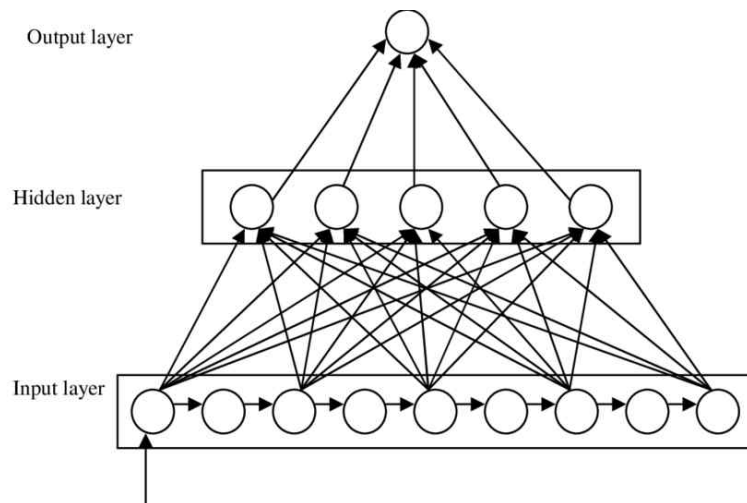
1.13.1.1 모델 학습

- 음향모델 설명

- 음향모델 학습에 사용하는 데이터는 음성과 라벨링 된 텍스트로, 음성에서 MFCC feature를 추출하고, 음성에 대응되는 음소를 TDNN과 TDNN-F라는 딥러닝 네트워크로 모델링하여 이 음소들의 변화를 HMM으로 모델링을 진행한다.
- 음성과 텍스트 쌍으로 하여 모델 학습을 완료하면, 음성 입력이 들어왔을 때 음성을 MFCC feature로 변화하여 DNN-HMM 모델에 적용하여 적절한 텍스트를 출력 진행한다.

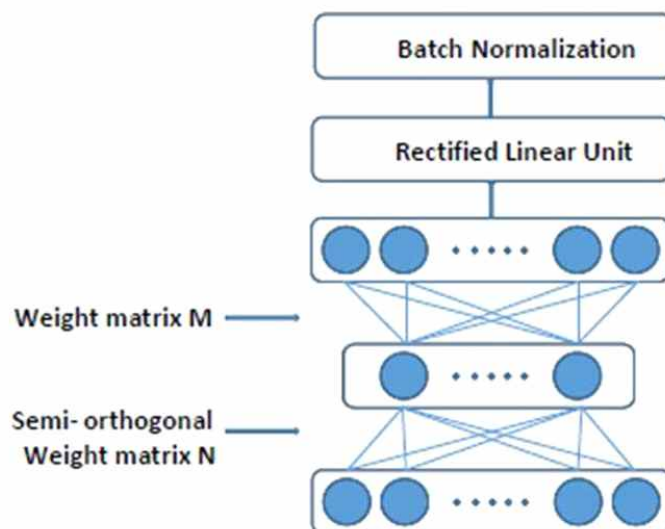
○ TDNN

다층 퍼셉트론(Multi Layer Perceptron) 구조를 가진 모델로써, 입력(input) 데이터로 사용하는 시간에 대해 현시점의 시간 값(t), 한 시점 전의 시간 값($t-1$), 두 시점 전의 시간 값($t-2$)을 한 번에 넣는 방법론이다. TDNN은 한정된 데이터의 time sequence만 학습할 수 있으므로 가장 중요한 long term dependency를 학습하기 어렵다는 문제가 있지만, 시간 의존성이 있는 데이터에서는 높은 학습 능력을 보여 유용하게 활용할 수 있다.



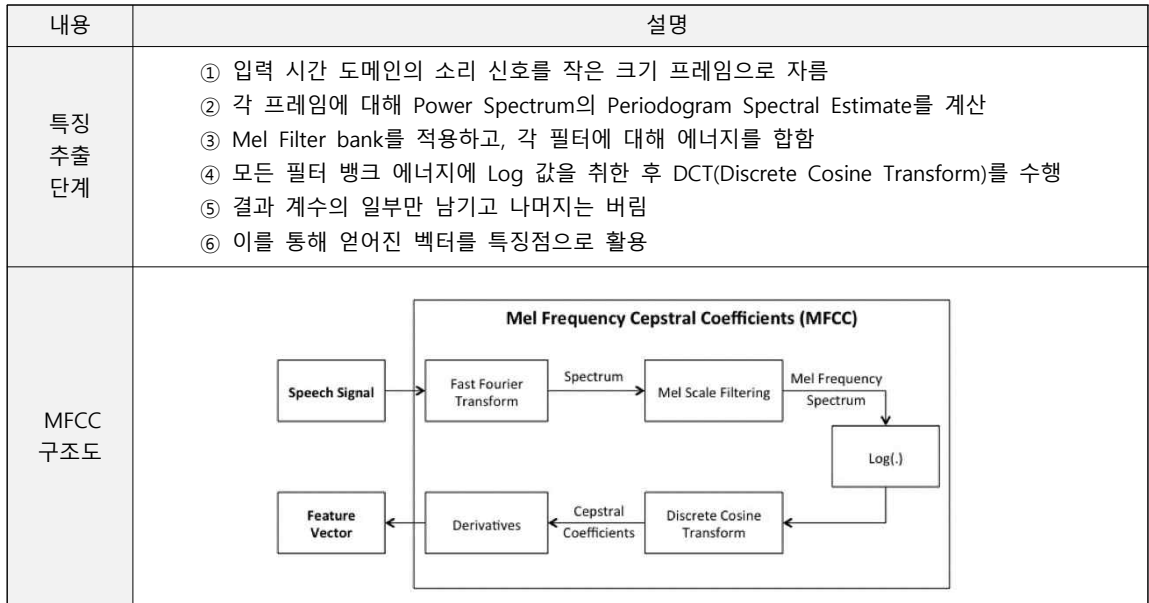
○ TDNN-F

Acoustic Model을 학습하는 네트워크 모델 중 하나로 1-d CNN이라고 볼 수 있는 기존 TDNN 에 Factorization을 추가한 방법이다. 해당 모델은 기존 TDNN Layer에 semi-orthogonal weight matrix를 넣어 factorize된 layer로 바꾼 것을 특징으로 한다.



○ MFCC

MFCC란 입력된 소리를 일정 시간 구간으로 나누어 각 구간에 대한 주파수 스펙트럼을 분석하여 특징을 추출하는 기법이다. 아날로그 형태의 음성 파형으로부터 수식적인 비교 연산이 가능하도록 특징점들을 추출(Feature Extraction)하는 과정이 필수적이다. 음성의 경우 통상적으로 MFCC (Mel-Frequency Cepstral Coefficients) 특징점으로 활용한다. MFCC는 다음과 같은 단계로 이루어져 있다.



○ HMM

음성신호가 입력되면 먼저 확률론적 기법을 통해 자동 인식이 시작된다. 이에 은닉 마르코프 모델 (Hidden Markov Model, HMM)이 사용된다. HMM은 가장 작은 단위의 음성인 음소(phone) 기반의 음향 모델을 구성하는 통계적 모델이다. HMM을 통해 음소와 음소 간의 전이 확률을 표시하는 방법으로 기초적인 음향 모델을 만들 수 있다. 일반적으로 음성인식에서는 음소를 기준으로 앞과 뒤의 음소와의 context에 따라 음소를 모델링하며 이렇게 3개의 음소로 모델링하는 것을 Tri-phone이라고 한다.

- 언어모델 설명

- 언어모델용 데이터는 방대한 양의 텍스트로 N-gram 모델을 생성하는 데 사용하며 음성을 텍스트로 바꾸는 디코딩 과정에서 음향모델 점수와 합해져 언어적 특성을 더 반영해주는 역할을 담당한다.
- Ngram은 언어의 최소 단위를 형태소 혹은 단어라고 할 때, 이 형태소 혹은 단어들이 연속적으로 나타나는 순서에 대한 확률분포이다. 이는 텍스트들은 서로 독립적이지 않고 언어에 따라 주어지는 문법, 습관에 따라 어떤 확률적인 관계를 맺고 있는 점에 착안해 만들어지는 모델이다. 언어모델은 텍스트가 주어졌을 때 형태소/단어의 순열에 대한 확률 정보를 바탕으로, 음성 모델이 판단한 형태소/단어의 순열을 입력받아 그 후에 이어질 텍스트를 판단함. 언어모델은 음성인식 디코딩 과정에서 음성 모델의 판단으로부터 텍스트 출력을 만들 때 참조되는 모델이다.
- 음성인식 기술에서는 전통적으로 N-gram 모델을 사용하여 언어모델을 구성한다. N-gram 모델은 대표적 확률 모델로서, n개 단어의 연쇄를 확률적으로 표현하면 언어 안에서 발생할 수 있는 모든 형태소/단어의 순서를 통계학적으로 모델링할 수 있다는 점에 착안해, 1개, 2개, ..., n개의 형태소/단어의 연속 발생 순서로 확률분포를 나타내는 N-gram 언어모델을 만들 수 있다. 일반 텍스트를 입력하면 이를 토대로 통계를 수행하여 N-gram 모델을 만들어주는 도구가 활용된다.

wood pittsburgh cindy jean
jean wood

일반 텍스트로부터
N-gram 모델 자동 생성

```

\data\
ngram 1=7
ngram 2=7

\1-grams:
-1.0000 <unk> -0.2553
-98.9366 <s> -0.3064
-1.0000 </s> 0.0000
-0.6990 wood -0.2553
-0.6990 cindy -0.2553
-0.6990 pittsburgh -0.2553
-0.6990 jean -0.1973

\2-grams:
-0.2553 <unk> wood
-0.2553 <s> <unk>
-0.2553 wood pittsburgh
-0.2553 cindy jean
-0.2553 pittsburgh cindy
-0.5563 jean </s>
-0.5563 jean wood

\end\
  
```

- 학습적용 방법
- 사전 요구사항

요구사항	세부 내용
서버 요구사항	<ul style="list-style-type: none"> - Nvidia Driver (>=450.80.02) - 해당 버전 드라이버 지원되는 1개 이상의 GPU - 도커 이미지 (nvidia/cuda:11.1-base-ubuntu20.04) 와 호환되는 OS
Docker	-
NVIDIA Docker	-
Docker 이미지 로드	<pre>\$ docker load -i nia_kaldi_image-22.01.26.NIA.tar</pre>

- 결과물 목록
- 도커 이미지

```

- image/
  └─ nia_kaldi_image-22.01.26.NIA.tar

```

- 모델

```

- model/
  └─ tdnn/
    ├── data/
    ├── exp/
    └── conf/
  └─ tdnnf/
    ├── data/
    ├── exp/
    └── conf/

```

- 학습 데이터

```

- train/
  └─ tmp/

```

- 테스트셋

```

- test/
  └─ testset/
    ├── 상/
    ├── 중/
    └── 하/

```

- 데이터 전처리용 스크립트

```

- scripts/
  └─ nia_foreigner_preprocessing.py

```

○ 학습진행 방법

- 충분한 학습량 확보를 위해 학습, 검증, 평가 데이터셋의 비율은 90:5:5로 한다.
- 결과물인 데이터 전처리용 스크립트를 사용해 Kaldi용 학습 데이터셋 {kaldi_train} 생성한다.

```

$ python nia_foreigner_preprocessing.py \
  {원본데이터}/ \
  {kaldi_train}/ \
  {testset}/ \
  {train_corpus}

```

- {원본데이터}는 아래와 같은 구조로 돼 있음을 가정한다.

```

{원본데이터}
└─ 9. 최종(납품본)
   └─ 1. 원천데이터
      └─ 1. 베트남어
         └─ 2. 영어
            └─ 3. 일본어
               └─ 4. 중국어
                  └─ 5. 태국어
                     └─ 6. 기타
                        └─ 2. 라벨링데이터
                           └─ 1. 베트남어
                              └─ 2. 영어
                                 └─ 3. 일본어
                                    └─ 4. 중국어
                                       └─ 5. 태국어
                                          └─ 6. 기타

```

- {kaldi_train}과 결과물 3), 4)를 학습용 폴더 {train_dir} 으로 복사한다.

```

{train_dir}/
└─ {kaldi_train}/

└─ tmp/
   └─ testset/

```

- {train_corpus} 로 한국어 N-gram Language Model을생성한다.

```

tmp/data/local/lm/
└─ daglo_lexicon
   └─ daglo_lm_fg.arpa.gz
      └─ daglo_lm_tg.arpa.gz
         └─ daglo_lm_tgmed.arpa.gz
            └─ daglo_lm_tgsmall.arpa.gz
               └─ daglo_morfessorseg

```

- tmp/ 내에 학습된 언어모델 샘플이 있다.
- zeroth 프로젝트의 한국어 N-gram LM 생성 스크립트를 사용한다.
(<https://github.com/goodatlas/zeroth/blob/master/s5/data/local/lm>)
- 도커 컨테이너를 실행한다.

```
$ nvidia-docker run -it -e LC_ALL=C.UTF-8 --ipc=host --volume {train_dir}:/mnt
--gpus all --shm-size=1g --ulimit memlock=-1 --ulimit stack=67108864
nia_kaldi_image:22.01.26.NIA
```

- 볼륨 마운트된 모델, 테스트셋 경로 symbolic link 지정한다,

```
$ mkdir /data && cd /data && ln -s /mnt/{train_dir}/tmp tmp
$ ln -s /mnt/{train_dir}/trainset FOREIGNER_UTTERANCE_DATASET_final
$ mkdir /test && cd /test && ln -s /mnt/testset/
FOREIGNER_UTTERANCE_DATASET_final_TESTSET
```

- 학습 및 테스트 실행한다.

```
$ cd /kaldi/egs/NIA/s0/ && ./run.sh
```

- 학습 시 nvidia driver 는 exclusive mode 로 설정한다.

```
$ nvidia-smi -c 3
```

- 학습 파라미터는 각자 서버 환경에 맞게 설정한다.

```
- run.sh
- nCPU
- local/chain/run_tdnn.sh, local/chain/run_tdnf.sh
- num_jobs_initial
- num_jobs_final
- num_epochs
- minibatch_size
- initial_effective_lrate
- final_effective_lrate
```

○ 테스트 진행 방법

- 학습 결과물인 모델과 테스트셋을 테스트 폴더 {test_dir}로 복사한다.

```
{test_dir}/
├─ tdnn/
├─ tdnf/
└─ testset/
```

- 도커 컨테이너를 실행한다.

```
$ nvidia-docker run -it -e LC_ALL=C.UTF-8 --ipc=host --volume {test_dir}:/mnt
--gpus all --shm-size=1g --ulimit memlock=-1 --ulimit stack=67108864
nia_kaldi_image:22.01.26.NIA
```


- 볼륨 마운트된 모델과 테스트셋 symbolic link를 지정한다.

```
$ mkdir /data && cd /data && ln -s /mnt/{tdnn or tdnnf}/ tmp
$ mkdir /test && cd /test && ln -s /mnt/testset/ testset
```

- 테스트를 실행한다.

```
$ cd /kaldi/egs/NIA/s0/
$ python3 test.py cer /test/testset # CER 체크
$ python3 test.py wer /test/testset # WER 체크
```

- 테스트 결과

CN11RC001_CN0016_20210730.wav	0.1786	0.4167	0.3214	0.6667	0.0357	0.0833
CN11RC002_CN0031_20210726.wav	0.0526	0.2353	0.0263	0.0588	0.2105	0.5294
CN11RC002_CN0079_20210730.wav	0.1842	0.3529	0.2105	0.2941	0.1053	0.2941
CN11RC002_CN0111_20210727.wav	0.2895	0.4118	0.1316	0.2941	0.1842	0.2353
CN11RC002_CN0211_20210726.wav	0.0	0.0	0.1842	0.3529	0.0789	0.2353
CN11RC004_CN0051_20210727.wav	0.0	0.0	0.2105	0.5	0.2895	0.6429
CN11RC005_CN0016_20210730.wav	0.12	0.2353	0.18	0.3529	0.08	0.1765
CN11RC005_CN0058_20210726.wav	0.0	0.0	0.18	0.4118	0.04	0.1765
CN11RC006_CN0036_20210831.wav	0.1163	0.3125	0.9535	1.0	0.0465	0.3125
CN11RC006_CN0203_20210726.wav	0.0	0.0	0.9535	1.0	0.0	0.25
CN11RC007_CN0021_20210730.wav	0.3488	0.5625	0.6047	0.75	0.0698	0.375
CN11RC007_CN0078_20210727.wav	0.0698	0.1875	0.5581	0.75	0.1628	0.5
CN11RC007_CN0111_20210727.wav	0.3721	0.75	0.6512	0.75	0.0465	0.3125
CN11RC007_CN0131_20210727.wav	0.0233	0.0625	0.5349	0.625	0.0465	0.3125
CN11RC007_CN0203_20210726.wav	0.093	0.25	0.6047	0.875	0.0698	0.3125
CN11RC007_CN0211_20210726.wav	0.1163	0.1875	0.5349	0.6875	0.093	0.3125
CN11RC007_CN0246_20210727.wav	0.0	0.0	0.5814	0.75	0.093	0.4375
CN11RC007_CN0259_20210730.wav	0.0698	0.1875	0.6279	0.75	0.0465	0.3125
CN11RC008_CN0178_20210822.wav	0.0351	0.0952	0.8947	0.9524	0.0877	0.2857
CN11RC009_CN0014_20210730.wav	0.0417	0.125	0.1042	0.1875	0.0833	0.375
CN11RC010_CN0002_20210730.wav	0.4167	0.6316	0.9375	1.0	0.4792	0.7368
CN11RC011_CN0024_20210730.wav	0.0455	0.1176	0.3864	0.5882	0.0	0.0
CN11RC016_CN0041_20210730.wav	0.0526	0.1333	0.1579	0.3333	0.1842	0.3333
CN11RC016_CN0046_20210730.wav	0.0263	0.0667	0.1842	0.3333	0.0789	0.5333
CN11RC016_CN0053_20210730.wav	0.0789	0.2	0.2105	0.4667	0.1053	0.5333
CN11RC016_CN0092_20210730.wav	0.1053	0.2667	0.3158	0.4667	0.2895	0.5333
CN11RC016_CN0102_20210730.wav	0.1053	0.2	0.1579	0.2667	0.2895	0.5333
CN11RC016_CN0109_20210730.wav	0.1579	0.2667	0.2368	0.4667	0.2368	0.6
CN11RC016_CN0134_20210727.wav	0.0263	0.2	0.2105	0.6667	0.1053	0.4667
CN11RC016_CN0195_20210727.wav	0.0526	0.1333	0.1842	0.3333	0.0	0.1333
CN11RC016_CN0207_20210730.wav	0.1053	0.2	0.2105	0.3333	0.1842	0.4
CN11RC016_CN0234_20210727.wav	0.0526	0.1333	0.0526	0.2	0.1053	0.4667

오류율	CER	WER
한국어 문장 구사 실력	Error Rate(%)	Error Rate(%)
상	8.88	24.47
중	9.90	26.51
하	9.61	26.38
Average Error Rate(%)	9.32	25.36

1.13.1.2 서비스 활용 시나리오(예시)

- 국내 거주 외국인을 위한 음성인식 스마트스피커
 - 한국어 발음이 서툰 외국인의 발음을 쉽게 인식하는 음성인식 엔진을 탑재한 스마트 스피커
 - 외국인의 한국어 발화 인식 정확도를 향상하여 외국인 또한 한국어 기반 음성인식 명령을 쉽게 내릴 수 있도록 지원한다.
- 기대효과
 - 증가하는 국내 거주하는 외국인이 편리한 생활을 할 수 있도록 지원할 수 있다.
 - 한국어 발화량을 늘리면서 한국어 학습에 도움을 줄 수 있다.
- 서비스 아이디어 내용
 - 정확한 한국어 발음만 인식하는 스마트스피커에 외국인 발화를 학습한 음성인식 엔진을 탑재하여 외국인 발화도 쉽게 인식할 수 있도록 하여, 인공지능 기술 친숙도를 높일 수 있다.



1.13.2 데이터 제공

외국인 발화 한국어 음성 데이터는 한국정보화진흥원이 운영하는 AI Hub를 통해 데이터와 데이터에 대한 메타데이터, 데이터 구축 가이드라인, 사용 및 설치 매뉴얼을 포함하는 저작도구, 기타 구축 관련 내용을 공개한다. 해당 데이터는 AI Hub 계정이 있는 누구나 사용할 수 있게 공개되며, AI Hub에서 다운로드 받을 수 있다.

1.13.3 데이터 유지보수

AI Hub 내 데이터 공개 시 구축량, 데이터 구축 내용, 데이터, 데이터 클래스 등을 함께 기재하여 추후에도 해당 데이터 사용과 활용이 용이할 수 있도록 한다. 데이터에 대한 문의는 AI Hub의 문의 게시판을 통해 문의할 수 있으며, 한국정보화진흥원 담당자를 거쳐 수행기관 담당자에게 전달된다. 전달된 문의에 대해서 개선 조치할 예정이다.