●○ 한국어 방언 과제

한국어 방언 발화 데이터(제주도)

●○ 개요: 한국어방언 AI 학습 데이터 셋이란?

모든 디지털 산업의 기초가 될 데이터는 80% 이상이 텍스트, 음성, 영상 등으로 되어 있고, 이중 음성, 텍스트 데이터는 인공지능 A.I를 학습시키기 위한 기술인 NLP(Natural Language Processing)의 핵심적인 부분을 차지하고 있습니다.

한국어 방언 AI 학습데이터 셋은 지역 간 언어적 특성 및 차별을 허물고 모두가 함께 배우고 활용할 수 있는 사람 중심의 AI 시대를 구현할 수 있는 학습 데이터 셋으로 주관사 ㈜솔트룩스를 포함한 15개 참여기업이 제주도 방언 3,000시간의 음성 데이터 셋과 50 만건의 전사 텍스트를 구축하였습니다.

표준어 및 방언을 넘어선 소통 체제와 이를 기반으로 한 각 분야별 활용 가능 지능형플랫폼 구축 및 AI 돌봄 서비스, 스마트시티 데이터 허브, 스마트팜 등 방언 음성 데이터 적용이 가능한 산업분야 및 실생활 AI 서비스 개발에 활용할 수 있습니다.

●○ 데이터셋의 구성

한국어 방언 AI 학습데이터 셋 구성에 대한 요약은 아래를 참고할 수 있습니다.

과제명	주요 내용	데이터 구축량	데이터 형식
한국어 방언 발화 데이터 (제주도)	방언(제주도)을 사용하는 일상 대화를 수집하여 음성을 문자로 변환한 방언 발화 데이터셋 구축	• 조용한 환경에서 ***2,000명*** 이상의 화자가 발화한 ***3,000시간*** 이상의 음성 데이터셋 • 원본 표준어 텍스트 및 방언 특성을 고려하여 전사한 텍스트 ***50만건***	• 원본형태 : 화자가 구분된 담화 텍스트 말뭉치 • 학습용 데이터 형태 : 방언 발화된 음성 데이터가 맵핑된 텍스트와 음성 데이터셋

데이터 종류	포함 내용	제공 방식
음성 데이터셋	총 3,000 시간 정제된 음성데이터	wav 포맷 파일
텍스트 데이터셋	총 50건의 원본 표준어 텍스트 및 방언 특성을 고려한 이중전사 텍스트	JSON 포맷 파일

●○ 데이터셋의 설계 기준과 분포

제주도 방언 일상대화 음성과 음성인식 기술을 활용해 텍스트로 변환한 텍스트 데이터셋으로 1) 연구분야: 음성 발화, 음성 인식, NLU, NLG를 포함한 NLP 전분야, 2) 산업분야: 온라인 심리상담, 고객상담 챗봇, 스마트 스피커 등에 활용할 수 있도록 방언 데이터가 맵핑된 이중전사 형식의 텍스트와 음성 데이터셋 형태입니다.

학습데이터 구축을 위한 구성 원칙과 주요 특징은 다음과 같다.

- 음성 녹음 화자의 구성
  - 특정 주제에 내용이 편중되지 않도록 사전 협의하여 진행
  - 한 화자당 최대 녹음시간은 가능한 약 30분으로 하고 동일 화자가 중복 참여하지 않도록 제한하나, 동일 주제가 아닐 경우에는 허용
  - 녹음 화자 모집 시 최초 2인 1조로 신청자를 최우선으로 하며, 1인이 개별 신청했을 경우 비슷한 연령대 및 관심사를 구분하여 조 편성
  - 주제에 따라 1인 녹음, 3인 이상 녹음을 허용

화자별 분류방법	세부 내용
연령별	1그룹(10대~20대), 2그룹(30대), 3그룹(40대 이상)
지역별	제주도 지역

  - 총 3개의 그룹으로 구성되며 1그룹은 10대~20대, 2그룹은 30대, 3그룹은 40대 이상으로 정하였으며 실제 녹음이 불가능하다고 판단되는 0~9세/70세 이상의 대상자는 제외이나 녹음이 가능할 경우 3그룹에 포함하여 진행
- 음성 데이터 전사 규칙

분류	전사규칙
개요	• 발화된 그대로 전사하는 발음 전사와 한글 맞춤법 및 표준어 규정에 따른 철자 전사를 병행하여 전사하는 것을 기본 원칙으로 한다.
화자 표시	• 화자 성별, 연령 등 화자 정보를 표시한다. 화자에 대한 정보를 모를 경우에는 '?'로 표시한다. • 본문 전사에서 화자 정보와 화자 표시는 반드시 일치해야 하고 화자가 분명하지 않을 경우에는 '?'로 표시한다.
전사 단위	• 기본 전사 단위는 긴 휴지, 경계 억양, 경계말 장음화 등을 특징으로 하는 억양구(IP: Intonational Phrase)가 되도록 한다.
숫자/외래어/기호/단위	• 한글 맞춤법, 표준어 규정, 외래어 표기법 등 관련 어문 규정에 따라 한글로 적는다.
발음	• 표준 발음에서 벗어난 형식으로 발화하거나 표준 발음이 여러 개인 경우 등 한글 맞춤법 표기에 따른 발음과 차이가 있는 경우에 발음 전사를 한다.
발화 겹침	• 겹침 발화는 표시하지 않고 시간 순서에 따라 적는다. • 만약 맞장구 발화가 일어날 경우 맞장구 발화를 사이에 넣어 주 발화를 나눈다.

분류	전사규칙
익명성 보장	• 대화자들의 신분 보장을 위해 이름, 주민등록번호, 카드 번호, 전화 번호 등 개인 정보와 관련된 사항은 노출되지 않도록 전사 단계에서 비식별화한다. - n : 사람 이름(단, 정치인, 연예인 등 유명인의 이름은 비식별화하지 않으며, 상호명은 부정적인 경우에만 비식별화) - social-security-num : 주민등록번호 - card-num : 신용카드 번호 - address : 주소(동 이하의 구체적인 주소만 비식별화) - tel-num : 전화 번호

- 제주도 방언 이중 전사 규칙
  - 방언 전사: 각 지역에서 모은 사람들의 대화를 지역 언어의 특성이 드러나도록 소리 나는 대로 적는 것.
  - 방법 및 유의 사항: 방언과 관련이 없는 표현은 표준어를 적는 방식으로 쓰되, 방언 표현은 방언의 특색을 드러나도록 표기한다. 이때 방언의 표기는 음성 그대로 소리나는 대로 쓰지 않고 방언의 형태가 드러나는 방식으로 쓴다.
  - "표준어 대응쌍 전사"는 소리 나는 대로 적은 "방언 전사"가 표준어 규정에서 벗어난 경우에, 그에 대응하는 표준형을 함께 제시하는 것을 원칙으로 한다. 띄어쓰기를 기준으로 하여 방언과 표준어를 각각 괄호 안에 넣어서 전사하고 이들 사이에는 빗금(/)을 넣는다. 방언 전사를 먼저하고 표준어 대응쌍 전사를 그 뒤에 나란히 제시

지역	보기
제주	아까 (집드레)/(집으로) (가라.)/(가더라.) 너 (하*구정 한)/(하고 싶은) 대로 (하*라.)/(해라.) (아매나)/(아무렇게나)

지역	보기	
	올바른 전사 표기	잘못된 전사 표기
제주	하루에 같이 (검질멧주게.)/(김매었지).	하루에 같이 (검질멧쭈게.)/(김매었지).

지역	보기
제주	성격이 참 (요망지다.)/(야무지다.) (하르방)/(할아버지) 댁에 가는 길.

●○ 데이터 구조

데이터셋에 따른 항목과 해당 값은 아래 테이블과 같다.

수준 1	수준 2	수준 3	수준 4	타입	설명
id				string	AI 학습데이터 파일 아이디
metadata				object	AI 학습데이터 파일 메타 정보
	title			string	AI 학습데이터 파일 제목
	creator			string	구축자: 솔트룩스
	distributor			string	배포자: 솔트룩스
	year			string	구축 년도: 2020
	category			string	분류: 구어 〉 사적 대화 〉 일상 대화
	annotation_level			array (string)	분석 층위: 원시
	sampling			string	샘플링 방식: 본문 전체
document				array (object)	대화 정보
	id			string	대화 아이디
	metadata			object	대화 메타 정보
		title		string	대화 제목: 2인 일상 대화
		author		string	저작권자: 개인 발화자
		publisher		string	발행자: 개인 발화 녹음
		date		string	녹음일자: YYYYMMDD
		topic		string	대화 주제
		speaker		array (object)	화자 정보
			id	string	화자 아이디
			age	string	연령
			occupation	string	직업
			sex	string	성별
			birthplace	string	출생지
			principal_residence	string	주 성장지
			current_residence	string	현 거주지
			education	string	학력
		setting		object	환경 정보
			relation	string	화자 간 관계
	utterance			array (object)	발화 정보
		id		string	발화 아이디
		form		string	방언 전사
		standard_form		string	표준어 대응쌍 부착
		speaker_id		string	화자 아이디
		start		num	발화 시작 시간

수준 1	수준 2	수준 3	수준 4	타입	설명
		end		num	발화 종료 시간
		note		string	전사자 기타 메모
		dialecteojeol		array (object)	방언 어절 단위 정보
			id	num	방언 어절 번호
			form	string	방언 어형
			begin	num	어절의 시작
			end	num	어절의 끝
		standardeojeol		array (object)	표준어 대응쌍 정보
			id	num	표준어 대응쌍 번호
			dialecteojeol_form	string	방언 어형(어절 단위)
			equivalent	string	표준어 대응쌍
			begin	num	어절의 시작
			end	num	어절의 끝
			dialecteojeol_id	num	방언 어절 번호
			position	num	표준어 대응쌍의 어절 내 위치 (1어절이 2어절 이상으로 나뉘는 경우에 필요함)

●○ 데이터 예시

이 데이터는 설명 가능 데이터 기준이며, 아래 예시와 같은 구조를 가진다.


```
{
"id": "SDRW2000000001",
"metadata": {
"title": " 경상방언 AI 학습데이터 SDRW2000000001",
"creator": "솔트룩스",
"distributor": "솔트룩스",
"year": "2020",
"category": "경상방언 〉 사적 대화 〉 일상 대화",
"annotation_level": [
"원시"
],
"sampling": "본문 전체"
},
"document": [
{
"id": "SDRW2000000001.1",
"metadata": {
"title": "2인 일상 대화",
```

```
"author": "개인 발화자",
"publisher": "개인 발화 녹음",
"date": "20190711",
"topic": "자동차",
"speaker": [
{
"id": "SD1900011",
"age": "30대",
"occupation": "사무 종사자",
"sex": "남성",
"birthplace": "대구",
"pricipal_residence": "대구",
"current_residence": "경북",
"education": "대졸"
},
{
"id": "SD1900012",
"age": "30대",
"occupation": "사무 종사자",
"sex": "남성",
"birthplace": "대구",
"pricipal_residence": "대구",
"current_residence": "대구",
"education": "대졸"
}
],
"setting": {
"relation": "동료"
}
},
"utterance": [
{
"id": "SDRW2000000001.1.1.1",
"form": "안녕하세요.",
"original_form": "안녕하세요.",
"speaker_id": "SD1900011",
"start": 30.56600,
"end": 32.48262,
"note": ""
},
{
"id": "SDRW2000000001.1.1.2",
"form": "아~ xx님 오랜만입니다.",
"original_form": "아~ ((xx님)) 오랜만입니다.",
"speaker_id": "SD1900012",
```

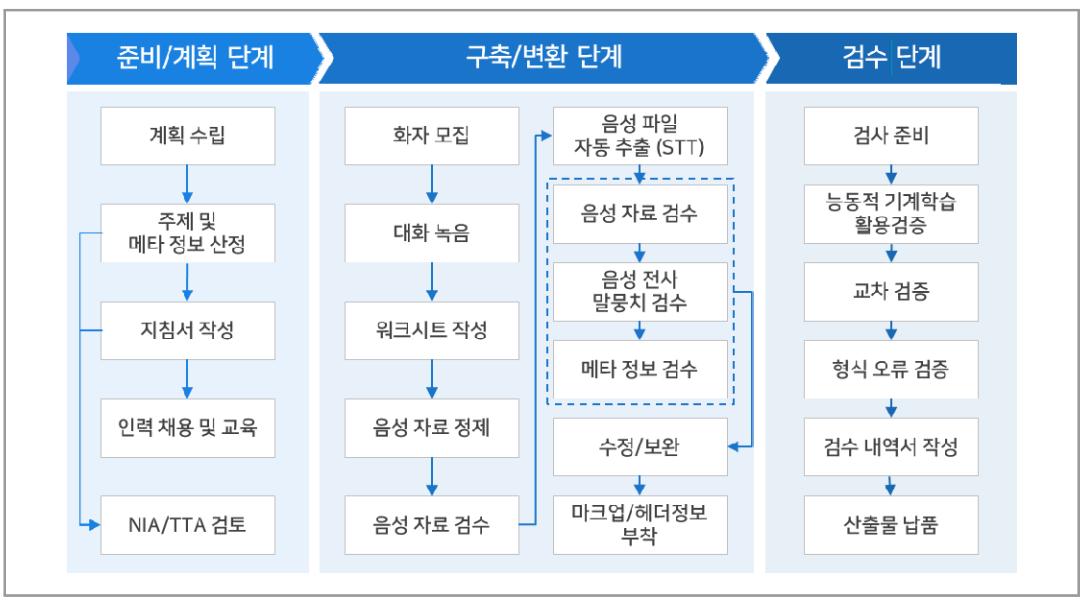
```
"start": 33.12500,
"end": 34.1543323,
"note": ""
},
{
        "id": "SDRW2000000001.1.1.3",
        "form": "은젬 예점보다늠 많이 단촐해졌다, 그지예?",
        "standard_form": "인제 예전보다는 많이 단촐해졌다, 그렇지요?",
        "speaker_id": "SD1900011",
        "start": 34.1543324,
        "end": 45.1543323,
        "note": "",
        "dialecteojeol": [
                {
                        "id": 1,
                        "form": "은젬",
                        "begin": 0,
                        "end": 2
                },
                {
                        "id": 2,
                        "form": "예점보다늠",
                        "begin": 3,
                        "end": 8
                },
                {
                        "id": 3,
                        "form": "많이",
                        "begin": 9,
                        "end": 11
                },
                {
                        "id": 4,
                        "form": "단촐해졌다,",
                        "begin": 12,
                        "end": 18
                },
                {
                        "id": 5,
                        "form": "그지예?",
                        "begin": 19,
                        "end": 23
                },
        ],
        "standardeojeol": [
```

```
                {
                        "id": 1,
                        "dialecteojeol_form": "은젬",
                        "equivalent": "언제",
                        "begin": 0,
                        "end": 2,
                        "dialecteojeol_id": 1,
                        "position": 1
                },
                {
                        "id": 2,
                        "dialecteojeol_form": "예점보다늠",
                        "begin": 3,
                        "end": 8,
                        "equivalent": "예전보다는",
                        "dialecteojeol_id": 2,
                        "position": 1
                },
                {
                        "id": 3,
                        "dialecteojeol_form": "그지예?",
                        "equivalent": "그렇지요?",
                        "begin": 19,
                        "end": 23,
                        "dialecteojeol_id": 5
                        "position": 1
                }
        ]
}
{
        "id": "SDRW2000000001.1.1.4",
        "form": "니 뭐라캤노?",
        "standard_form": "너 뭐라고 했니?",
        "speaker_id": "SD1900011",
        "start": 46.1543323,
        "end": 48.1543323,
        "note": "",
        "dialecteojeol": [
                {
                        "id": 1,
                        "form": "니",
                        "begin": 0,
                        "end": 1
                },
                {
```

```
                        "id": 2,
                        "form": "뭐라캤노?",
                        "begin": 3,
                        "end": 7
                }
        ],
        "standardeojeol": [
                {
                        "id": 1,
                        "dialecteojeol_form": 니",
                        "equivalent": "너",
                        "begin": 1,
                        "end": 1,
                        "dialecteojeol_id": 1,
                        "position": 1
                },
                {
                        "id": 2,
                        "form": 뭐라캤노?",
                        "equivalent": "뭐라고",
                        "begin": 3,
                        "end": 7,
                        "dialecteojeol_id": 2,
                        "position": 1
                },
                {
                        "id": 3,
                        "form": 뭐라캤노?",
                        "equivalent": "했니",
                        "begin": 3,
                        "end": 7,
                        "dialecteojeol_id": 2,
                        "position": 2
                }
        ]
}
```

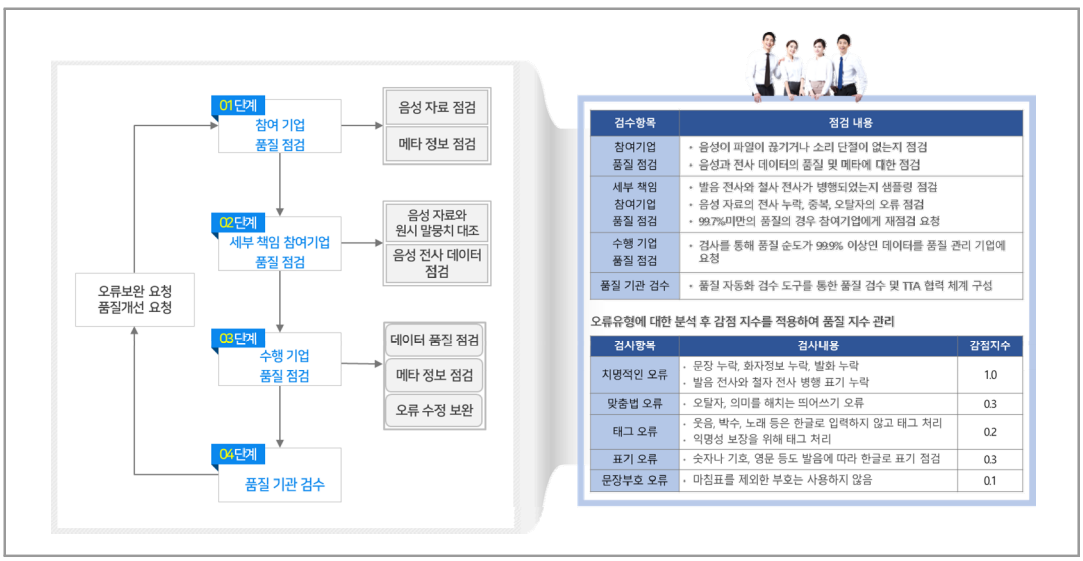
●○ 데이터 구축 과정

한국정보화진흥원의 데이터베이스 구축방법론(Ver.4)을 적용하여 음성 녹음, 이중 전사, 원시 말뭉치 구축에 대한 대상 자료별 공정 태스크와 주요 활동 절차를 표준화하여 효율적인 학습용 데이터셋 구축 체계를 확보하고 한국어 방언 AI데이터 구축에 적합하도록 자료의 특성을 고려하여 준비/계획 단계, 구축/변환단계, 검수단계의 3단계 공정을 설계한다.



●○ 검수와 품질 확보

데이터셋에서는 4단계 검수 체계를 구축하였으며, 가장 하위 레벨에는 크라우드 워커들이 작업한 결과물을 전사규칙 가이드라인에서 제시한 형식에 맞는지 체크하는 참여기업에 1차 검수자가 있으며, 이들이 검수한 결과물에 대해서 재검수하는 도별 책임기관 2차 검수자가 있습니다. 이렇게 만들어진 데이터셋을 전체적으로 들여다보며 데이터셋의 밸런스나 가이드라인의 적절성 등 품질 확보를 위한 총괄기업인 ㈜솔트룩스에서 5년 이상 학습데이터 구축 및 품질검수를 수행한 지식큐레이션 팀이 3차 검수를 수행하여 음성 및 전사 텍스트에 대한 품질을 확보하며, 최종 4차 검수는 참여기관인 ㈜비투앤에서 정확도 및 유효성 품질 검증을 통하여 고품질 학습 데이터셋의 품질을 담보할 수 있었습니다.



검수항목	점검 내용
참여기업 품질 점검	• 음성이 파열이 끊기거나 소리 단절이 없는지 점검 • 음성과 전사 데이터의 품질 및 메타에 대한 점검
세부 책임 참여기업 품질 점검	• 발음 전사와 철사 전사가 병행되었는지 샘플링 점검 • 음성 자료의 전사 누락, 중복, 오탈자의 오류 점검 • 99.7%미만의 품질의 경우 참여기업에게 재점검 요청
수행 기업 품질 점검	• 검사를 통해 품질 순도가 99.9% 이상인 데이터를 품질 관리 기업에 요청
품질 기관 검수	• 품질 자동화 검수 도구를 통한 품질 검수 및 TTA 협력 체계 구성

오류유형에 대한 분석 후 감점 지수를 적용하여 품질 지수 관리

검사항목	검사내용	감점지수
치명적인 오류	· 문장 누락, 화자정보 누락, 발화 누락 · 발음 전사와 철자 전사 병행 표기 누락	1.0
맞춤법 오류	· 오탈자, 의미를 해치는 띄어쓰기 오류	0.3
태그 오류	· 웃음, 박수, 노래 등은 한글로 입력하지 않고 태그 처리 · 익명성 보장을 위해 태그 처리	0.2
표기 오류	· 숫자나 기호, 영문 등도 발음에 따라 한글로 표기 점검	0.3
문장부호 오류	· 마침표를 제외한 부호는 사용하지 않음	0.1

●○ 데이터 구축 담당자

수행기관(주관) : ㈜솔트룩스
(전화: 02-2193-1674, 이메일: jwpark@saltlux.com)