●○ 상황별음성 과제

상담 음성 데이터

●○ 개요: 상담음성 데이터란?

- 본 과제의 데이터 구축 목적은 AI 상담센터를 위한 음성상담 음성인식기술 및 언어이해, 언어생성 연구 및 서비스 개발에 활용할 목적으로 한다. 한국인의 음성을 문자로 바꾸어 주고, 문맥을 이해하는 한국어 음성언어처리 기술 개발을 위한 AI 학습용 한국어 음성 DB 구축을 목표로 유무선, 웹 기반 등 다양한 방식으로 상담센터에 연락하여 상담하는 내용을 녹음한 음성 데이터를 구축한다.
- 본 가이드라인으로 구축한 상담음성 데이터는 낮이든, 밤이든, 24시간, 365일 운영 가능한 AI 콜센터 구축을 위한 음성인식 모델 생성과 다양한 대화 시나리오 구축을 위한 기본 대화 데이터 셋으로 활용 가능함.
- 또한 기존의 상담사와 AI 상담사를 함께 운영하는 하이브리드 콜센터로 상담사의 피로도와 업무 스트레스를 줄여주어 대고객 서비스를 더욱 성실히 수행할 수 있도록 함.
- 이러한 AI 콜센터를 통해 고객은 상담 대기로 긴 시간을 소비하지 않고, 상담을 원하시는 시간에 언제든 금융/구매/정보조회 등의 서비스를 제공 받을 수 있음.
- 본 상담음성 데이터의 사례에 대해서는 아래의 그림을 참고할 수 있다.

●○ 데이터셋 구성

원천데이터 종류	수집 채널	수집 시간
교육 도메인	크라우드 소싱 녹취 (가상시나리오 기반)	1000시간
금융 도메인	크라우드 소싱 녹취 (가상시나리오 기반)	1000시간
통신판매 도메인	크라우드 소싱 녹취 (가상시나리오 기반)	1000시간
총 구축량		3,000시간

- 본 데이터셋은 교육, 금융, 통신판매 도메인을 기준으로 3,000시간으로 구성되어 있음
- 문장별로 전사된 시간 기준으로 데이터량 산정
- 전문장 앞뒤의 0.5초 이내의 묵음 시간이 포함된 Speech Signal로 데이터 구축량 산정
- Speech Signal 만으로 3000 시간 이상의 데이터 구축
- 상담음성 AI 학습용 데이터 구축량은 다음과 같다.

구축 시간	발화자 수	도메인 수
3,000시간 이상	1,000명 이상	3개

- 구축되는 상담 데이터는 특정 도메인에 국한되지 않고, 다양성을 확보할 수 있도록 2개 이상의 도메인으로 데이터셋을 구성했으며, 음성인식 성능을 높일 수 있도록 1000명 이상의 녹취 인원을 확보했다.
- 데이터셋의 음원 파일은 압축 변환되지 않은 아래의 원본 파일로 구축

파일 형식	코덱	샘플레이트	채널
WAV	PCM	8K	Mono

●○ 데이터셋의 설계 기준과 분포

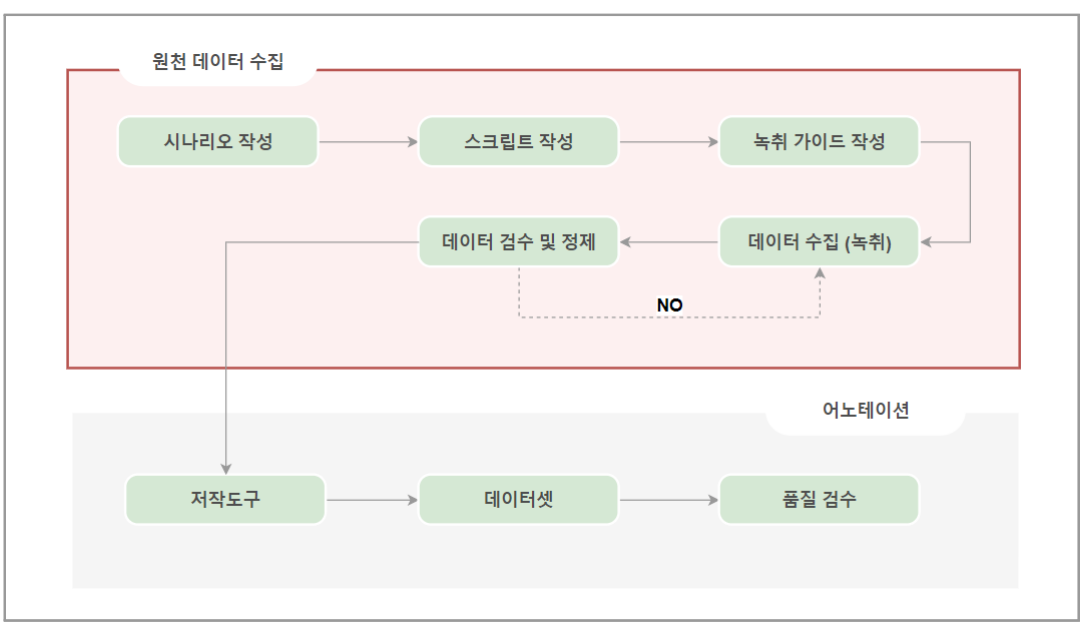
- 투자를 받지 못한 스타트업 기업은 다양한 고객 대응 서비스를 위한 콜센터를 제공하기 위한 비용을 부담스러워하며, 본업무 외에 자신의 휴대폰으로 주문, 상담, 환불, 배송조회 등의 콜센터 업무까지 직접 수행함.
- 또는, 콜센터 운용을 하고 있지만 서비스의 특성상 야간에서 고객 문의에 대응해야 하고 다양한 대 고객 서비스가 필요한 경우, 일반 콜센터의 업무 시간이 18시로 제한되어 있어서 적극적인 대 고객 서비스를 제공하지 못해 매출 신장의 한계와 소비자의 불만이 쌓여감.
- 이러한 문제점을 해결하기 위해서는 낮이든, 밤이든, 24시간, 365일 대 고객 서비스를 제공할

수 있는 AI 상담센터의 필요성이 대두되며, 이러한 AI 콜센터 구축을 위한 데이터셋을 제공하는 것을 목적으로 데이터셋을 설계 하였음

- 다양한 업체들을 위한 AI 상담센터를 위한 음성인식 학습데이터를 구축하기 위해서는 다양한 서비스 도메인으로 구성한 데이터셋 구성이 필요함.

도메인 종류	설명
교육 도메인	전용 학습기 기반의 온라인 교육 서비스에 대한 구매/환불/상담 등의 콜센터 상담 데이터
금융 도메인	보험/은행 등의 상담/계좌개설/분실신고 등의 콜센터 상담 데이터
통신판매 도메인	통신판매 등의 구매/정보문의/상담 등의 콜센터 상담 데이터

- 원천데이터 수집 절차는 데이터 시나리오 및 스크립트 작성, 녹취 가이드 작성, 크라우드소싱을 통한 데이터 녹취, 데이터 정제 단계로 이루어진다.



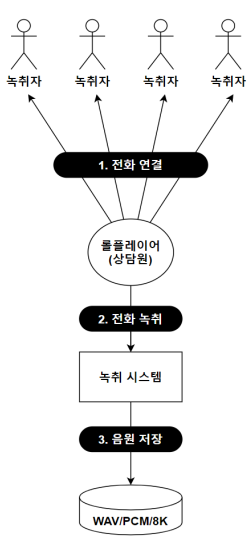
- 고객상담 시나리오는 상담사와 고객간의 긴 대화(3~10분)를 작성해야 하는 이슈로 시나리오 데이터 수집 방안을 잘 마련해야 하며, 본 과제에서는 콜센터의 전임 상담사들을 통해 원청사의 정보와 고객의 정보를 식별할 수 없도록 작성된 시나리오 데이터를 수집한다. 구축된 데이터를 활용도를 높이기 위해서는 다양한 도메인의 상담 시나리오 데이터를 확보하는 것이 중요하다.
- 녹취는 고객 부분에 대해서만 녹취 작업이 이루어지지만 자연스러운 녹취가 가능하도록 상담사의 대화 문장도 시나리오에 함께 포함하여 작성해야 한다.

도메인	분류	화자	문장
금융	대출문의	상담사	안녕하세요. 과제은행입니다.
		고객	신규 대출을 받고 싶어서요.

- 다양한 소스를 바탕으로 작성된 시나리오에 대해 검수 인력을 통해 검수 작업을 진행한다.

검수 항목	설명
도메인 확인	• 도메인 분류가 각 업종 그룹별로 맞게 정리가 되었는지 확인한다.
분류 확인	• 너무 세부적으로 나누어지지 않았는지 확인하여 지정된 카테고리 목록 안에서 설정되도록 조정한다.
화자 확인	• 문장에 맞는 화자가 설정되어 있는지 확인한다.
문장 확인	• 작성된 문장이 고객의 대화인지 확인한다. • 작성된 문장이 구어체인지 확인하고, 문어체이면 구어체로 변환한다. • 문장들의 대화 내용의 분류에 맞게 잘 구성되었는지 확인한다. • 작성된 문장에 개인정보식별 데이터가 포함되어 있으면 수정한다. - 이름: 대표적인 가명으로 변환 - 전화번호: 불가능한 조합의 번호 (010-0123-2344) - 주소: 동까지만 작성

- 상담데이터는 전화상의 녹취 작업이 이루어져야 하며 자연스러운 발화가 될 수 있도록 환경 구성이 이루어져야 한다. 본 과제에서는 전화 녹취와 자연스러운 발화를 유도할 수 있도록 콜센터 업체와 계약을 통해 기존의 상담 시스템을 활용한 전화 녹취와 상담원과의 전화 통화를 통해 자연스러운 대화를 유도할 수 있는 환경을 제공한다.



① 롤플레이어 역할을 맡은 상담사가 지정 크라우드워커에게 전화를 연결하여 상담 시나리오에 의한 고객 역할의 상담 녹취를 진행한다.
② 크라우드워커에 고객 역할의 상담 녹취는 콜센터의 녹취 시스템에 의한 녹취된다.
③ 녹취된 음원 파일은 본 과제의 음원 형식에 맞도록 WAV, PCM(코덱) 형태의 데이터로 수집 저장 관리한다.

- 데이터 수집에서는 아래와 같은 내용을 주로 고려하였다.

고려사항	설명	비고
서비스와 활용성	• 시나리오 작성 시, 다양한 서비스 분야에 적용될 수 있도록 2개 이상의 도메인 • 실제 데이터셋 활용을 고려하여 추가 데이터 등록을 최소화 하도록 슬롯 엔티티의 아이템 다양화	
소음	• 실제 활용 환경을 고려한 환경 소음 포함	
데이터의 균형	• 녹취되는 성별의 균형 • 상담 카테고리의 다양성 • 한 시나리오에 대한 중복 녹취 50명 이하	

●○ 데이터 구조

- 데이터셋에 따른 항목과 해당 값은 테이블과 같다.

No	항목		타입
	한글명	영문명	
	데이타셋	dataSet	
1	데이터셋 버전	version	String
2	녹취된 음원의 URL	mediaUrl	String
3	녹취된 날짜	date	String
4	음원 데이터 상세 정보	typeInfo	
4-1	음원 카테고리 정보 : 강의, 회의, 고객응대 등	category	String
4-2	음원 서브카테고리	subcategory	String
4-3	음원 녹취 장소	place	String
4-4	화자 목록	speakers	List
4-3-1	화자 유형 : 강사, 상담사, 고객, 기타	type	String
4-3-2	인입 유형 : 유선, 모바일, 인터넷 등	gender	String
4-5	화자 성별 : 남성, 여성	inputType	String
5	전사 데이터 목록 : 화자가 변경될 때마다 생성	dialogs	List
5-1	화자 아이디 : speakers 에 등록된 순번	speaker	String
5-2	전사된 텍스트	text	String
5-3	전사된 텍스트의 음원 재생 시작 위치	startTime	String
5-4	전사된 텍스트의 음원 재생 끝 위치	endTime	String
5-5	전사된 텍스트 문장과 관련된 태그 리스트	tags	String

●○ 데이터 예시

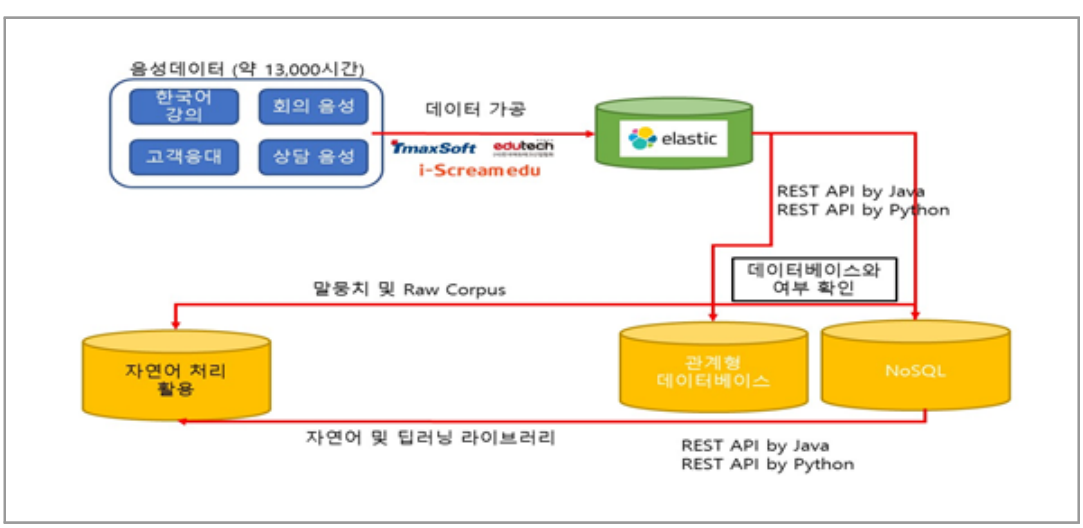

```
{
        "dataSet":
        {
                "version": "1.0",
                "mediaUrl": "http://..../23skskjdsfsks.wav",
                "date": "2020/05/20",
                "typeInfo":
                {
                        "category": "conference"
                        "speakers":
                        [
                                {
                                        "type": "representative", "gender": "female"
                                },
                                {
                                        "type": "customer", "gender": "female"
                                }
                        ],
                        "inputType": "mobile"
                "dialogues":
                [
                        {
                                "speaker": 0,
                                "text": "안녕하세요, 고객님. 무엇을 도와드릴까요?",
                                "startTime": "10.212",
                                "endTime": "12.432",
                                "tags": [ "Question", "Intro" ]
                        },
                        {
                                "speaker": 1,
                                "text": "환불할려고요",
                                "startTime": "12.569",
                                "endTime": "13.698",
                                "tags": [ "Answer", "Refund" ]
                        }
                ]
        }
}
```

●○ 데이터 구축 과정

- AI 콜센터 등의 서비스를 위한 학습용 데이터는 데이터 수집, 데이터 정제, 데이터 가공, 데이터셋 검수의 4단계를 거쳐 구축된다.

구축 단계	세부 절차	설명
수집	수집 도메인 선정	구축하려고 하는 서비스 업종 등의 적용 분야를 선정함.
	서비스 시나리오 작성	제공하려는 서비스 시나리오를 수립하여 작성함.
	시나리오별 스크립트 작성	하나의 시나리오에서 다양한 경우의 문장들을 작성하고 상품명 같은 엔티티를 다양하게 적용하여 작성함.
	크라우드소싱을 이용한 녹취	콜센터의 특징상 전화 녹취가 될 수 있는 환경을 제시해야 하며, 동일 문장의 녹취 인원이 최대 50명이 넘지 않도록 함.
정제	원천 데이터 검수	스크립트와 다른 녹취, 과도한 소음이 포함, 녹취자의 목소리 너무 낮은 음원 등을 제거하는 과정을 통해 데이터를 정제
가공	가공 인력 교육	인공지능 학습용 데이터 가공에 필요한 작업 교육과 훈련 수행
	데이터 가공	음원의 녹취와 스크립트 상의 불일치 부분에 대해 스크립트 수정과 음원을 문장 단위로 분리 저장.
검수	음성 모델 학습	구축된 데이터셋의 음원으로 음성모델 생성
	전수 검수	생성된 음성 모델을 통해 자동 전사된 텍스트와 스크립트와 텍스트 비교

●○ 검수와 품질확보



- 음성데이터를 고려한 품질관리 프로세스 마련
  - 음성 데이터를 전사하여 가공한 데이터를 통해 자연어 연구에 활용할 수 있는 목적을 기준으로 품질 검수 항목을 선정하고 프로젝트 진행간 지속적으로 현실에 맞도록 추가 및 수정
- 데이터 활용에 중점을 둔 품질관리 프로세스 마련
  - 가공 데이터를 자연어 연구에 쉽게 적용할 수 있는 지표 확인
  - 자연어 연구자가 쉽게 최신 트렌드에 맞는 라이브러리 및 애플리케이션에 적용할 수 있는지 여부를 확인
  - 13,000여 시간에 대한 데이터를 사용자가 쉽게 사용할 수 있는 크기 및 파일명 선정
- 가공 데이터의 2차 검수시 자동화
  - 정성적인 아닌 정량적인 검증 지표를 만들고, 자동화하여 3번이상의 크로스체크 및 이력관리
- 정확도 · 유효성 검증 계획서 기준의 외부전문기관 (TTA) 검증 수행
  - 품질의 핵심은 사용자가 쉽고 유익하게 사용하는 것에서 나온다는 기본 전제아래 철저히 검증 수행하여 최종적인 데이터셋의 품질을 담보할 수 있었다.