

메타데이터 정보 (다중기입가능)	분야	데이터 유형	구축 데이터량	원천데이터 형식	라벨링 형식	라벨링 유형
	<i>N/A 기입</i>	오디오	3,000시간	wav	json	자유대화음성 (자연어)
	데이터 출처	데이터 구축년도	구축기관(총괄)	가공기관	검수기관	
	자체 수집	<i>2022년</i>	미디어젠(주)	(주)비디	(주)메트릭스 미디어젠(주)	
	데이터 문의처	기관명	문의담당자명	전화번호 (유선전화번호기입)	메일주소	
		미디어젠(주)	송민규	02-6429-7104	minks@media zen.co.kr	
	데이터 소개	감정이 태깅된 청소년의 대본이 없는 자유대화 음성 데이터				
	주요키워드	한국어, 자유대화, 음성, 일상대화				
카테고리 정의서		첨부의 카테고리 정의서 엑셀파일에 데이터카테고리 작성하여 제출(예시참고)				

데이터셋명	국 민 감 정 대 사	감정이 태깅된 자유대화 데이터																	
	Emotion-tagged free conversation audio data																		
구축목적	청소년의 자유대화를 수집하되 대본이 없이 진행되는 일상 대화 음성 데이터를 구축 음성 데이터 전사 정보, 감정 상태 및 유형 정보 등이 태깅된 데이터 구축																		
활용서비스	자연어처리 모델 개발, 대화형 인공지능 시스템 연구, 대화형 로봇/앱 개발, 가상의 아바타 응답 패턴 개발 등																		
소개	별도로 개발된 APP을 통해 2 사람이 다양한 주제에 대해 실시간 음성 통화를 하며, 해당 통화 내역을 텍스트로 전사함 각각의 음성은 감정 상태, 감정 유형, 감정 정도, 감정 대상 등으로 세분화된 감정 태깅을 수행함																		
데이터셋 통계 (구축 규모 및 분포)	1. 데이터 구축 규모 - 총 3,000시간의 음성 데이터 구축																		
	<table><tr><th rowspan="2">구분</th><th>총 데이터수집 (건)</th></tr><tr><th>청소년</th></tr><tr><td>데이터 수집 규모</td><td>3,000 시간</td></tr></table>		구분	총 데이터수집 (건)	청소년	데이터 수집 규모	3,000 시간												
	구분	총 데이터수집 (건)																	
		청소년																	
	데이터 수집 규모	3,000 시간																	
	2. 청소년 데이터 분포 (3,000시간) - 대화 참여자 성별 분포 (청소년)																		
	<table><tr><th>성별</th><th>인원(명)</th><th>비율(%)</th></tr><tr><td>남성</td><td>1,605</td><td>36.21%</td></tr><tr><td>여성</td><td>2,828</td><td>63.79%</td></tr><tr><td>합계</td><td>4,433</td><td>100%</td></tr></table>		성별	인원(명)	비율(%)	남성	1,605	36.21%	여성	2,828	63.79%	합계	4,433	100%					
	성별	인원(명)	비율(%)																
	남성	1,605	36.21%																
	여성	2,828	63.79%																
합계	4,433	100%																	
- 대화 참여자 연령대 분포 (청소년)																			
<table><tr><th>연령대</th><th>수량(명)</th><th>비율(%)</th></tr><tr><td>10대</td><td>4,433</td><td>100.00%</td></tr><tr><td>합계</td><td>4,433</td><td>100%</td></tr></table>		연령대	수량(명)	비율(%)	10대	4,433	100.00%	합계	4,433	100%									
연령대	수량(명)	비율(%)																	
10대	4,433	100.00%																	
합계	4,433	100%																	
- 대화 참여자 지역 분포 (청소년)																			
<table><tr><th>연령대</th><th>수량(명)</th><th>비율(%)</th></tr><tr><td>서울/인천/경기</td><td>3,507</td><td>79.11%</td></tr><tr><td>대전/세종/충청/강원</td><td>357</td><td>8.05%</td></tr><tr><td>광주/전라/제주</td><td>161</td><td>3.63%</td></tr><tr><td>부산/대구/울산/경상</td><td>408</td><td>9.20%</td></tr><tr><td>합계</td><td>4,433</td><td>100%</td></tr></table>		연령대	수량(명)	비율(%)	서울/인천/경기	3,507	79.11%	대전/세종/충청/강원	357	8.05%	광주/전라/제주	161	3.63%	부산/대구/울산/경상	408	9.20%	합계	4,433	100%
연령대	수량(명)	비율(%)																	
서울/인천/경기	3,507	79.11%																	
대전/세종/충청/강원	357	8.05%																	
광주/전라/제주	161	3.63%																	
부산/대구/울산/경상	408	9.20%																	
합계	4,433	100%																	
- 주제 분포: 교육/학문, 엔터테인먼트, 쇼핑, 게임 등 11개 주제 대분류 (청소년)																			

순서	주제	수량(건)	비율(%)
1	교육/학문	1,136	9.71%
2	엔터테인먼트	1,908	16.30%
3	쇼핑	800	6.84%
4	게임	611	5.22%
5	고민Q&A	2,280	19.48%
6	음식	1,140	9.74%
7	스포츠	388	3.32%
8	건강	438	3.74%
9	취미	1,770	15.13%
10	생활	1,231	10.52%
합계		11,702	100%

- 감정 상태 분포 (청소년)

순서	주제	수량(건)	비율(%)
1	기쁨	1,863,236	50.86%
2	놀라움	300,512	8.20%
3	두려움	68,425	1.87%
4	사랑스러움	180,509	4.93%
5	슬픔	192,043	5.24%
6	화남	212,975	5.81
7	없음	845,813	23.09%
합계		212,975	5.81%

- 감정 유형 분포 (청소년)

순서	주제	수량(건)	비율(%)
1	긍정	2,389,045	65.21%
2	부정	485,114	13.24%
3	중립	789,354	21.55%
합계		3,663,513	100%

데이터셋 구성

1. 데이터셋 구성

- 청소년 3,000시간의 음성 데이터 구축
- 2사람의 대화를 각각 녹음한 음성 파일(16kHz, 16bit, mono)이 하나의 대화 파일로서 음성 파일을 구성(16kHz, 16bit, stereo)

- 2사람이 각각 30분씩 대화를 한 경우, 두 모노 파일을 합쳐서 총 대화 시간 30분 스테레오로 저장
- 텍스트 전사는 각각의 화자에 대해 개별 진행하고, 하나의 Json 파일에 메타 정보와 함께 기재
- 7개 감정 상태(기쁨, 놀라움, 두려움, 사랑스러움, 슬픔, 화남, 없음)로 구분하여 태깅
- 그 외 감정 정도(강함, 보통, 약함), 감정 대상(특정 인물/사물, 특정 사건, 대화 태도), 감정 종류(긍정, 중립, 부정) 등 감정과 관련된 태깅 정보를 기재
- 각 음성 파일 당 1개의 Json 파일이 존재하며, 메타정보와 모든 대화 텍스트를 저장함

2. 데이터 분류 체계 정의

- 청소년 자유대화 : 감정이 태깅된 청소년 자유대화

3. 데이터 출처

- 크라우드 워커를 모집하여 자유대화 직접 구축

4. 라벨링 데이터 구성

구분	속성명	타입	필수여부	설명	범위
1	File	Object		파일 정보	
1-1	FileName	String	y	파일명	
	FileLength	String	y	파일 길이	
	BeginOfSpeech	String	y	발화 시작점	
	EndOfSpeech	String	y	발화 끝점	
	SpeechLength	String	y	발화 길이 (초)	
2	Wav	Object		원천데이터 정보	
2-1	SamplingRate	String	y	주파수	
	NumberOfBit	String	y	비트수	
	EncodingLaw	String	y	인코딩방식	
	NumberOfChannel	String	y	채널수	
3	Noise	Object		소음 정보	
3-1	Speaker1NoiseCategory	String	y	실내/실외 환경 종류	실내/실외
	Speaker1NoiseType	String	n	녹음 장소	
	Speaker1SNR	String	n	SNR	
	Speaker2NoiseCategory	String	y	실내/실외 환경 종류	실내/실외
	Speaker2NoiseType	String	n	녹음 장소	
	Speaker2SNR	String	n	SNR	
4	Speaker1	Object		참여자 정보	
4-1	ID	String	y	발화자 고유 번호	0001~9999
	AgeCategory	String	y	청소년 구분	청소년
	Age	String	y	연령	A1~A6
	Gender	String	y	성별	G1/G2
	Residence	String	n	거주지	
	Occupation	String	y	직업	
	Condition	String	n	목상태	
	Mask	String	n	마스크 착용 유/무	착용/미착용
	RecDevice	String	n	녹음 장치	

5	Speaker2	Object		참여자 정보	
5-1	ID	String	y	발화자 고유 번호	0001~9999
5-2	AgeCategory	String	y	청소년 구분	청소년
5-3	Age	String	y	연령	A1~A6
5-4	Gender	String	y	성별	G1/G2
5-5	Residence	String	y	거주지	
5-6	Occupation	String	y	직업	
5-7	Condition	String	n	목상태	
5-8	Mask	String	n	마스크 착용 유/무	착용/미착용
5-9	RecDevice	String	n	녹음 장치	
6	ConversationInfo	Object		대화정보	
6-1	Domain	String	y	대화 도메인	
6-2	Step1Subject	String	y	대화 토픽	
6-3	Step2Subject	String	y	대화 토픽	
7	Conversation	Object		대화내용	
7-1	TextNo	String	y	대화 순서 번호	000001~999 999
7-2	SpeakerNo	String	y	발화자 구분	Speaker1/ Speaker2
7-3	SpeakerEmotionCategory	String	y	감정 유형	긍정/부정/ 중립
7-4	SpeakerEmotionTarget	String	y	감정 상태	기쁨/사랑스 러움/두려움 /화남/슬픔/ 놀라움/없음
7-5	SpeakerEmotionLevel	String	y	감정 정도	약함/보통/ 강함
7-6	SpeakerEmotionObject	String	y	감정 대상	0/1/2
7-7	VerifyEmotionCategory	String	y	감정 유형	긍정/부정/ 중립
7-8	VerifyEmotionTarget	String	y	감정 상태	기쁨/사랑스 러움/두려움 /화남/슬픔/ 놀라움/없음
7-9	VerifyEmotionLevel	String	y	감정 정도	약함/보통/ 강함
7-10	VerifyEmotionObject	String	y	감정 대상	0/1/2
7-11	Text	String	y	Text 전사내용	
7-12	StartTime	String	y	대화 시작시간	
7-13	EndTime	String	y	대화 끝 시간	

<주요 상세 정보>

1. 유형 : 1 (청소년), 2 (성인)
2. 성별 : G1 (남성), G2 (여성)
3. 연령층 : A1 (10대) ~ A6 (60대 이상)
4. 노이즈 타입 : T1 (실내), T2 (실외)
5. 도메인 정보 : 1D01~1D10 (청소년 관련 도메인), 2D01~2D10 (성인 관련 도메인)
6. 대화 끝 번호 : ~C999999
7. 녹음 일련번호 : 중복값 방지를 위한 녹음 일련 번호

<파일명 정보 부가 설명>

1. 파일명 샘플 : 1_6690G2A1_7864G1A1_T1_1D05T0197C000001_035597 2. 파일명 샘플에 대한 설명 : 청소년_1번화자ID+여성+10대_2번화자ID+남성+10대_실내_도메인+토픽+대화끝번호_중복방지일련번호						
데이터셋 구축 수행기관 담당자	주관기관	기관명	책임자명	전화번호 (유선전화번호)	메일주소	담당업무
		미디어젠(주)	송민규	02-6429-7104	minks@mediazen.co.kr	과제 총괄
	참여기관	기관명	담당업무	기관명	담당업무	
		(주)비디	수집 및 검수 시스템 개발			
		(주)메트릭스	데이터 수집, 가공, 정제, 검수			