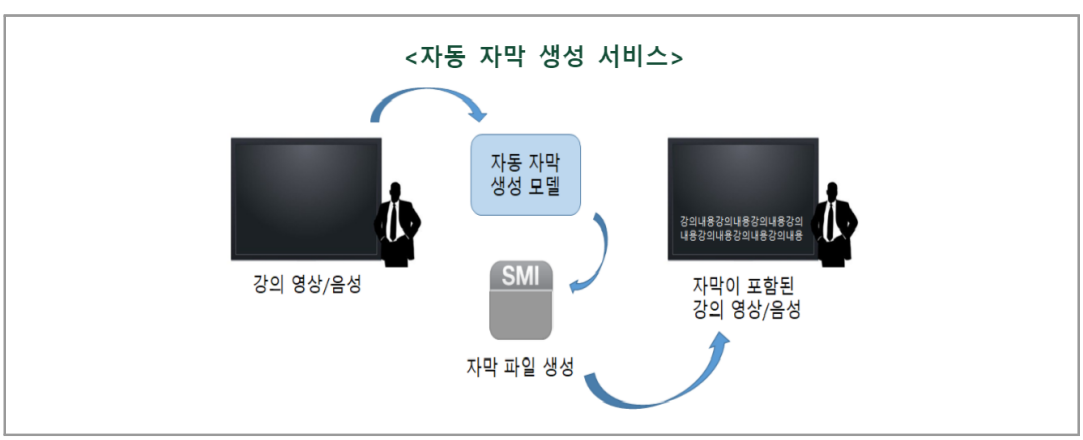
●○ 상황별음성 과제

한국어 강의 데이터

●○ 개요: 한국어 강의 데이터란?

- 한국인의 음성을 문자로 바꾸어 주고, 문맥을 이해하는 한국어 음성언어처리 기술 개발을 위한 AI 학습용 한국어 음성 DB를 구축.
- 한국어로 된 강의영상/음성을 인식하여 자동으로 자막을 생성해주고, 내용을 이해하는 서비스를 위한 한국어 강의 음성DB를 구축.
- AI 학습용 데이터 구축량 : 한국교육방송공사(EBS)로부터 제공받은 4,000 방송시간 이상 분량의 한국어 강의 데이터이며, 발화자의 남녀비율은 1:1로 구성, 1개 강의는 20~40분 내외로 구성.
- 사용자는 자막을 생성하고자 하는 강의 영상/음성을 자동 자막 생성기에 입력하면 입력에 대한 자막이 결과로 생성되어 저장.
- 결과 자막 파일에는 각 문장이 발화된 시간과 그에 대응하는 문장이 저장되며 영상 플레이어를 실행하여 강의 영상과 자막을 아래 그림과 같이 함께 시청.
- 사용하고자 하는 영상 플레이어의 플랫폼에 따라 자막 파일이 저장되는 포맷을 변경하여 사용할 수 있으며, 이는 차후 다양한 서비스 환경에서 유연성을 가짐.
- 본 서비스는 추후 실시간 음성인식이 가능한 모델을 도입하여 실시간으로 진행되는 강의에 자동으로 자막을 생성해주는 서비스로 발전시킬 수 있음.



●○ 데이터셋의 구성

- 초등학교, 중학교, 고등학교의 국어, 수학, 사회, 과학, 역사 5종 과목 및 직업/자격증, 기타 성인강의로 구성된 4,000시간의 한국어 강의 데이터로 구성되어 있으며, 모든 데이터 셋은 음성, 스크립트로 구성되어 있음.
- 문장별로 전사된 시간 기준으로 데이터량 산정.
- 전사된 문장의 시간에는 문장 앞뒤의 묵음 시간이 포함되지 않음으로 실제 Speech Signal만으로 데이터 구축량 산정 가능..
- Speech Signal 만으로 4000 시간 이상의 데이터 구축.

구분	초	중	고	직업/자격증	기타 성인
시간(hr)	1,250시간 이상	750시간 이상	1,200시간 이상	500	300
데이터 선정 기준	- 학년별 3단계 수준 분류 - 수준별 15시간 음성데이터 생성 총 12학년 (초, 중, 고) 국어, 수학, 사회, 과학, 역사 5종 15시간X12학년X3단계X5종 ÷ 2,000시간			- 직업/자격증 분야별 100시간 음성데이터 생성 - 공무원, 안전, 식품, 공인중개사 등 10종 10종X50시간=500시간	10종 분류 (인문, 철학, 문학, 예술, 과학, 사회, IT,교육) 종별 30시간
수집항목	- 강의음성, 강사정보(성별,이름), 학년, 과목, 수준, 강좌별 강의수, 강의시간 등			강의음성, 강사정보(성별, 이름), 강의주제, 강의수, 강의정보, 발화장소, 강의대상 등	

- 데이터의 음성 품질 기준

음성 품질 준수	- Extension: PCM - precision: 16-bit - Sample rate: 16k Hz - channel: mono - Sample Encoding: 16-bit Signed Integer PCM

●○ 데이터셋의 설계 기준과 분포

- 최근 AI를 활용한 음성인식 기술이 인공지능 비서를 포함한 다양한 서비스에 적용되고 있으나, AI 음성인식 모델이 한국어 강의 데이터로 학습이 이루어지지 않을 경우, 강의에서 발생하는 다양한 노이즈 등으로 인해 음성인식 성능이 떨어지는 경향을 보임.
- 기존에 공개된 한국어 강의 도메인에 대한 데이터셋은 존재하지 않으며, 저작권 문제 등으로 활용이 불가함에 따라, 본 과제에서 한국어 강의 도메인에 대한 AI 음성인식 성능 향상을 위한 한국교육방송공사(EBS)로부터 제공받은 4,000 시간 이상의 한국어 강의 데이터셋을 구축함.

- 한국어 강의 도메인은 넓은 범위를 다룰 필요가 있으며, 이를 위해 크게 3가지 도메인으로 나누었으며, 각 도메인의 세부 수집 대상은 아래와 같음.

  * 학생 (초중고): 국어, 수학, 사회, 과학, 역사

구분	초등학교						중학교			고등학교		
학년	1	2	3	4	5	6	1	2	3	1	2	3
국어	30	30	45	45	50	50	50	50	50	80	80	80
수학	30	30	45	45	50	50	50	50	50	80	80	80
사회	30	30	45	45	50	50	50	50	50	80	80	80
과학	30	30	45	45	50	50	50	50	50	80	80	80
역사	30	30	45	45	50	50	50	50	50	80	80	80
합계	150	150	225	225	250	250	250	250	250	400	400	400

  * 직업/자격증: 공무원, 안전, 공인중개사 등 10종

구분	국어	한국사	사회	과학	수학	전문 자격	금융	경영	IT	기술
계획 시간	50	50	50	50	50	50	50	50	50	50
합계	50	50	50	50	50	50	50	50	50	50

  * 기타: 인문, 철학, 문학, 예술, 과학, 사회, IT, 교육

구분	인문	철학	예술	사회	IT	교양	문학	과학	교육	기타
계획 시간	30	30	30	30	30	30	30	30	30	30
합계	30	30	30	30	30	30	30	30	30	30

- 다양한 화자 분포를 가질 수 있도록, 화자 별 음성 길이를 최대 10시간으로 제한함
- 구축한 강의 음성 데이터셋은 500명 이상의 발화자로 구성하며, 음성 별 발화자에 대한 메타데이터를 제공함.
- 녹음은 약 128㎡, 400㎡, 600㎡, 800㎡ 공간에서 슈어 제조사의 UR4D+ 모델 또는 제나이져 SK5212/SKM5200 모델을 사용하여 이루어졌으며, 발화자와 마이크 사이의 거리는 약 20cm 에서 50cm 사이임.
- 데이터 전사 과정에서 앞뒤의 묵음구간을 포함하도록 싱크를 조절하였으나, 일부 데이터의 경우 발화 사이의 묵음구간이 짧을 수 있음.
- 1,4절 음성 전사 규칙에 발음겹침, 잡음, 말더듬, 텍스트정규화/비정규화, 발화자 처리 등에 대한 처리 방안을 제시함.

- 발음겹침 음성의 경우 메인 발화자 이외의 발화자에 대한 텍스트 전사에 상당한 어려움이 있어, 메인 발화자만 텍스트 전사하고 해당 어절에 전사 규칙에 따라 '+'를 표시함.
- 잡음이 포함된 음성의 경우 전사자와 검수자의 주관에 따라 충분히 인식이 가능한 경우만 포함하였음.
- 데이터 전사 과정에서 앞뒤의 묵음구간을 포함하나, 일부 데이터의 경우 발화 사이의 묵음구간이 짧을 수 있음.
- 데이터셋 구축 과정에서 성명, 전화번호, 주소 등의 개인정보가 포함된 음성은 배제함.
- 성차별, 정치적 성향, 종교 등 사회적 민감 정보를 포함된 음성은 배제함.
- 공개된 한국어 강의 데이터셋은 적합한 컨텐츠 구매과정을 통하였으며, 저작권자의 허가를 받아 데이터 사용 및 활용이 자유로움.
- 데이터셋의 효용성을 측정하기 위해 구축 데이터 기반의 상용화 가능한 수준의 AI 응용 서비스 개발을 진행함.

●○ 데이터 구조

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


```
{
  "metadata": {
    "title": "LH3D02S0089",
    "creator": "조주연(hcjy97@naver.com)",
    "distributor": "",
    "year": "2020년",
    "category": "",
    "annotation_level": [],
    "sampling": ""
  },
  "speaker": [
    {
      "no": 1,
      "name": "M1",
      "age": null,
      "sex": "남",
      "shortcut": 1,
      "occupation": "선생님"
    }
  ],
  "setting": {
    "relation": ""
  },
  "utterance": [
    {
      "id": "fee2333d-d35a-4f1e-a3ba-2575b8486f9f",
      "start": 8,
      "end": 16.66,
      "speaker_id": 1,
      "form": "여러분 정말 반갑습니다. 세상의 빛이 될 여러분을 진심으로 응원합니다. 경향 곧판왕 정종영입니다.",
      "standard_form": "여러분 정말 반갑습니다. 세상의 빛이 될 여러분을 진심으로 응원합니다. 경향 곧판왕 정종영입니다.",
      "note": "",
      "eojeolList": [
        {
          "id": "c82f58a4-dce9-4fd1-a689-9630edb87d2b",
          "eojeol": "여러분",
          "standard": "여러분",
          "begin": 0,
          "end": 3,
          "isDialect": false
        },
        {
          "id": "507ee3d1-a4fd-4a92-a90e-7c9420f542a1",
          "eojeol": "정말",
          "standard": "정말",
          "begin": 4,
          "end": 6,
          "isDialect": false
        },
        {
          "id": "f1ccef45-6237-4940-bcd6-de9953441db8",
          "eojeol": "반갑습니다.",
          "standard": "반갑습니다.",
          "begin": 7,
          "end": 13,
          "isDialect": false
        },
        {
          "id": "3a1c1825-45d9-4f9b-ae72-c29ebf337b24",
          "eojeol": "세상의",
          "standard": "세상의",
          "begin": 14,
          "end": 17,
          "isDialect": false
        },
        {
          "id": "591b66ad-86e6-425c-9f88-4f99cf62d2ac",
          "eojeol": "빛이",
          "standard": "빛이",
          "begin": 18,
          "end": 20,
          "isDialect": false
        },
        {
          "id": "7cb5f48f-bedb-4ef6-bb2e-f67ba14b400e",
          "eojeol": "될",
```

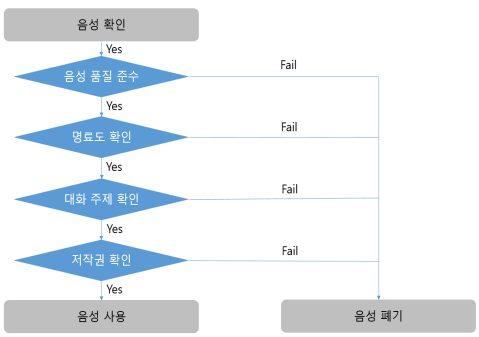

- 상기데이터는 전체 전사한 JSON 파일의 일부임

●○ 데이터 구축 과정

- 획득 정제 절차

원천 데이터 획득, 정제 절차

1. 음성 품질 준수 : 음성파일의 손실 여부 확인 (잡음, 파일 크기, 변환 파일 형식 등)
2. 명료도 확인 : 대화 음성 인지 여부 확인 (일반적으로 알아 들을 수 있는 정도 확인)
3. 대화 주제 확인 : 편향적인 주제의 여부 확인 (정치적, 특정 집단의 이익적인 측면 확인)
4. 저작권확인 : 법적인 측면 해결 유무 확인 (개인정보보호, 저작권 문제 해결)



- 획득 정제 기준

구분	절차	기준
획득	1. 음성 품질 준수	- Extension: PCM - Precision: 16-bit - Sample rate: 16kHz - Channel: mono - Sample Encoding: 16-bit Signed Integer PCM
정제	2. 명료도 확인	- 클리핑, Frame drop 등 손실된 음성데이터 제외 - 심한 잡음으로 사람도 인식하기 어려운 음성데이터 제외 - 비음성구간 제외
	3. 대화 주제 확인	- 민감한 이슈 (정치적 견해, 개인정보, 특정인물 비하, 성적인 표현) 발언이 포함된 경우 제외
	4. 저작권 확인	- 지적재산권 및 개인정보보호 관련 사항 해결 유무
	5. 최종 결정	- 4가지 절차에 모두 이상이 없는 경우

●○ 검수와 품질확보

- 데이터 검수는 양질의 데이터를 얻는 데 필요한 작업으로 정성적/정량적 평가를 통해 데이터의 유효함을 판별함
- 웹 저작도구를 사용하여 전사한 어노테이션 결과에 대해 1차, 2차, 3차 검수자를 걸쳐 음성품질 / 어노테이션 정확도 / 대화 주제 및 저작권확인 / 목표데이터 수집량 달성 여부를 확인함
- 1, 2차 검수자는 음성품질 및 어노테이션 정확도를 위주로 검수를 진행하며, 교차 검증을 통해 데이터의 오류를 최소화함
- 3차 검수자는 대화 주제 및 저작권확인, 목표데이터의 도메인별 수집량 달성 여부를 위주로 검수를 진행하며, 데이터의 다양성을 보장할 수 있도록 함
- 각 차수의 검수자는 피드백을 통해 데이터가 한쪽에 치우치지 않도록 하며, 전사자가 숙달할 수 있도록 지원함
- 3차 검수 완료 이후 TTA를 통한 데이터 검증 절차를 걸치며, 동시에 AI 모델 학습을 진행하여 데이터의 적합성과 유효성을 최종 판별함
- 어노테이션 결과 피드백 기준

대항목	피드백 목표	비고
음성품질	• 음성 파일의 샘플레이트, 채널, 인코딩의 적절성 여부 • 음성의 명료함 정도 (SNR)	전사자와 검수자 주관에 맡기며, 과반수 동의 시 적절함으로 판정
어노테이션 정확도	• 음성에 대응하는 텍스트 어노테이션 정확성 여부 • 음성에 대응하는 발화자 정보의 정확성 여부 • 음성과 텍스트의 싱크 정확성 여부	상기와 동일
대화 주제 및 저작권확인	• 대화 주제의 편향성 여부 • 저작권 및 개인정보 침해 여부	상기와 동일
목표데이터 수집량 달성	• 데이터 도메인, 화자 등 수집량의 적절성 여부	상기와 동일

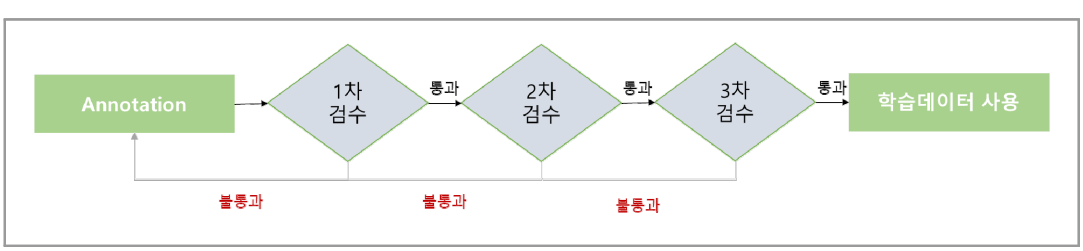


그림 | 어노테이션 결과 검증 절차

- 어노테이션 데이터 항목별 세부 판정 기준

대항목	데이터 항목	어노테이션	판정 기준
음성품질	음성 파일 형식	• N/A	• 전사도구 내 음성 파일 체크를 통한 자동 검수
	명료함	• N/A	• 전사자와 검수자에 의한 음성 명료함 판단
어노테이션 정확도	텍스트	• 전사규칙에 따라 전사된 발화 내용	• 3인 이상의 검수자 판단하에 과반수 통과
	발화자	• 화자 고유번호, 화자 성별, 나이	• 3인 이상의 검수자 판단하에 과반수 통과
	싱크	• 음성과 텍스트의 시작과 끝	• 3인 이상의 검수자 판단하에 과반수 통과
대화 주제 및 저작권 확인	대화주제	• 강의 전체 도메인에 대한 23종 태깅	• 3인 이상의 검수자 판단하에 과반수 통과
	저작권 및 개인정보	• N/A	• 3인 이상의 검수자에 의한 저작권 및 개인정보 침해여부 판단
목표데이터 수집량 달성	도메인	• 강의 도메인 별 음성 길이	• 음성 발화자 및 싱크 정보를 바탕으로 알고리즘을 통한 자동 계산
	화자	• 화자 별 음성 길이	• 음성 발화자 및 싱크 정보를 바탕으로 알고리즘을 통한 자동 계산

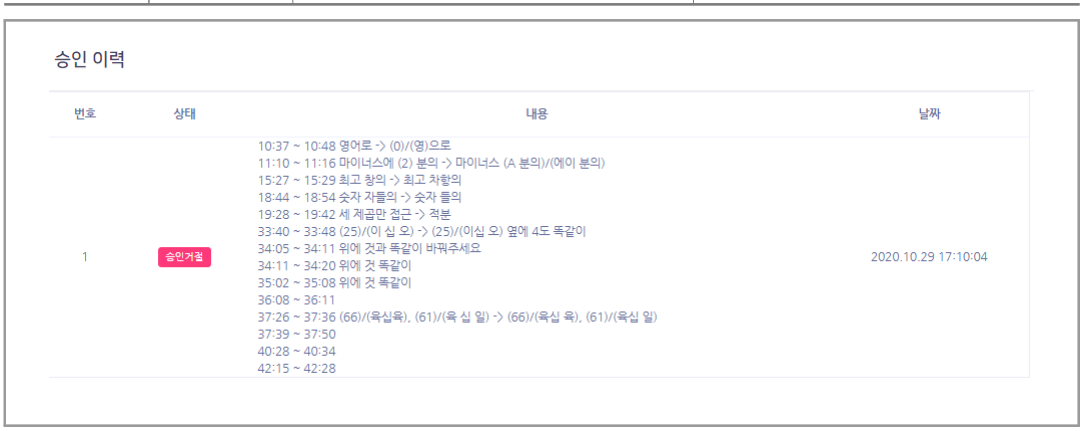


그림 | 검수자 피드백 예시