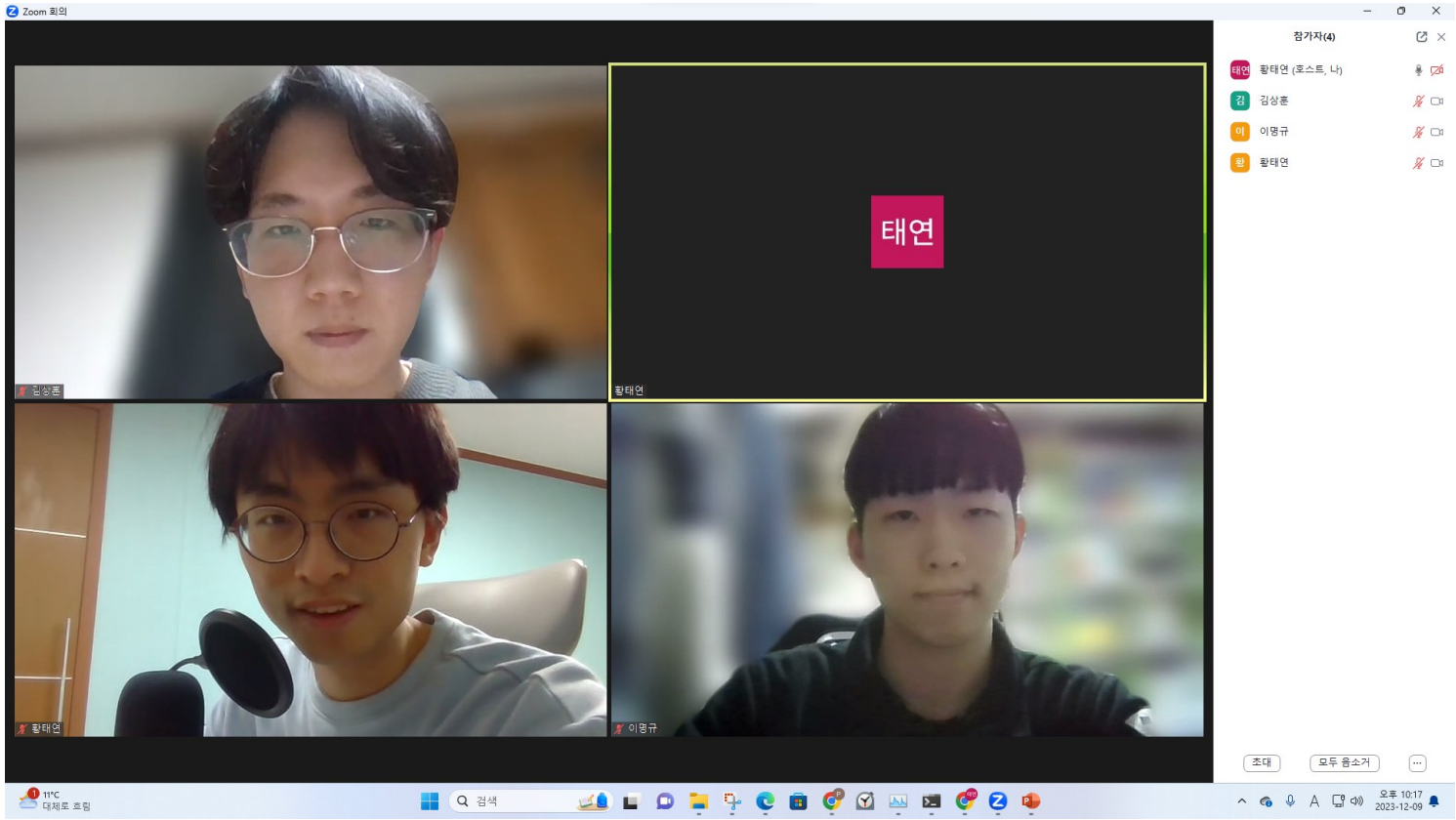
소프트웨어응용 프로젝트 최종 발표

가정 내 위급상황 음성 인식 모델 개발

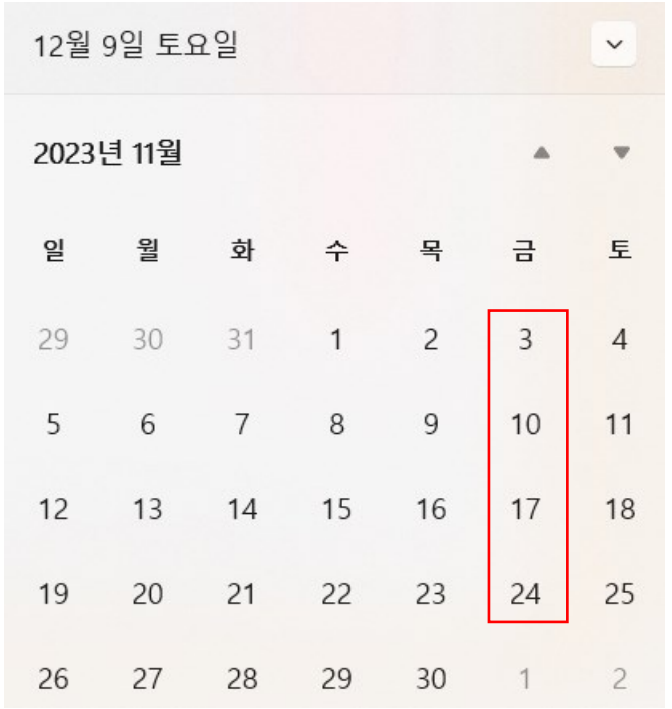
3조
2019540040 황태연　2019920014 김상훈　2020540023 이명규

프로젝트 회의 사진

3조 | 수학과 2019540040 황태연 | 컴퓨터과학부 2019920014 김상훈 | 수학과 2020540023 이명규 |



▲ 온라인 회의 단체 사진 (2023.12.09.)



▲ 매주 금요일 오후 2시 정기 회의 진행

1. 데이터 셋

- 데이터 셋: AI Hub 위급상황 음성/음향 데이터
  - 용량 : 508.04 GB

- 치안안전: 1. 강제추행, 2. 강도범죄, 3. 절도범죄, 4. 폭력범죄
- 소방안전: 5. 갇힘, 6. 전기사고, 7. 가스사고, 8. 화재, 9. 응급의료
- 자연재해: 10. 태풍/강풍, 11. 지진
- 사고 발생: 12. 낙상, 13. 붕괴사고
- 일반(위급): 14. 도움요청
- 일반(정상): 15. 실내, 16. 실외

- 대조군: 소음 환경 음성인식 데이터

17. 가전소음_세탁기, 건조기
18. 가전소음_청소기
19. 가전소음_기타소음

1. 데이터 셋

- 데이터 전처리

  - 학습 데이터 용량 : 60GB / 테스트 데이터 용량 : 11GB

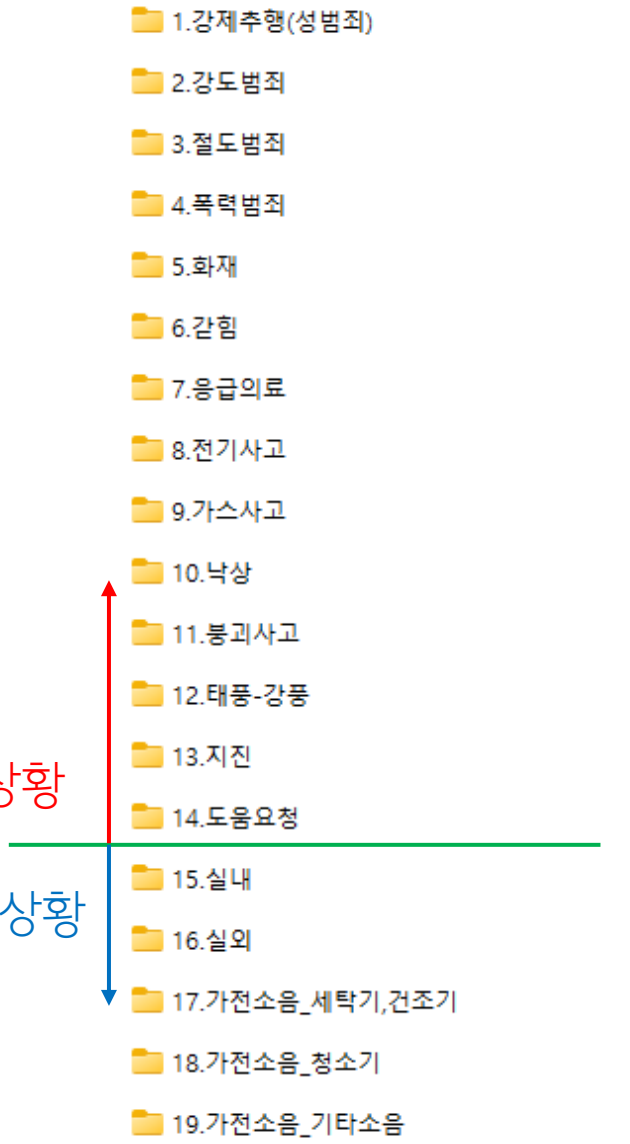
  - 각 분류당 학습 데이터 3000개 / 검증 데이터 1% / 테스트 데이터 500개

```
DatasetDict({
    train: Dataset({
        features: ['input_features', 'labels', 'class'],
        num_rows: 65848
    })
    test: Dataset({
        features: ['input_features', 'labels', 'class'],
        num_rows: 666
    })
})
```

▲ 학습/검증 데이터셋(Training/Validation Dataset) 구조

```
DatasetDict({
    test: Dataset({
        features: ['input_features', 'labels', 'class'],
        num_rows: 12102
    })
})
```

▲ 테스트 데이터셋(Test Dataset) 구조



2. 사용 시나리오

- 사용자가 우리의 어플리케이션을 어떻게 사용할 수 있을까?

- **1. 실내 위급상황에 대한 즉각적 대처**

1. 가정 내에 항상 음성을 인식할 수 있는 마이크를 설치

2. 마이크로부터 받아들인 음성 신호들을
   일상 생활 음성인지 위급 상황 음성인지 주기적으로 판별

3. 위급상황에 해당되는 음성으로 판별될 경우
   사용자에게 실제로 위급상황인지 물어보고

4. 10~30초간 응답이 없을 경우 위급 상황으로 판단하고,
   가족(또는 119)에게 음성 파일과 함께 위치 정보 및 상황 전달

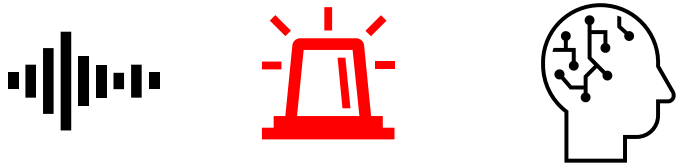
2. 사용 시나리오

- 사용자가 우리의 어플리케이션을 어떻게 사용할 수 있을까?

- 2. 특정 사고 다발 예상 지역에 위급상황 대처 인프라 구축

1. 특정 사고 다발 지역(ex 교통사고 다발 지역 등)에
   음성을 인식할 수 있는 기기를 설치

2. 기기로부터 받아들인 음성 신호들을
   해당 위급 상황 음성(ex. 교통사고, 붕괴, 화재)인지 주기적으로 판별



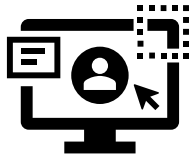
3. 위급상황에 해당되는 음성으로 판별될 경우
   소방서 및 경찰서에 음성 파일과 함께 위치 정보 및 상황 전달

2. 사용 시나리오

- 사용자가 우리의 어플리케이션을 어떻게 사용할 수 있을까?

- 3. 청각 장애인의 위험상황 인지 보조

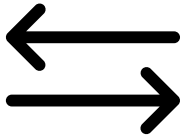
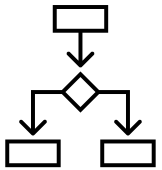
1. 청각 장애인이 해당 어플리케이션을 착용

2. 어플리케이션이 받아들인 음성 신호들을
   위급 상황 음성인지 주기적으로 판별



3. 위급상황에 해당되는 음성으로 판별될 경우
   청각 장애인에게 **시각적으로** 상황 전달 및 그 정보를 바탕으로
   청각 장애인의 신속한 대처 가능

3. 프로젝트 전체 구조 및 주요 내용

1. 기존의 Whisper 모델

- Encoder :

  셀프 어텐션 메커니즘을 통해

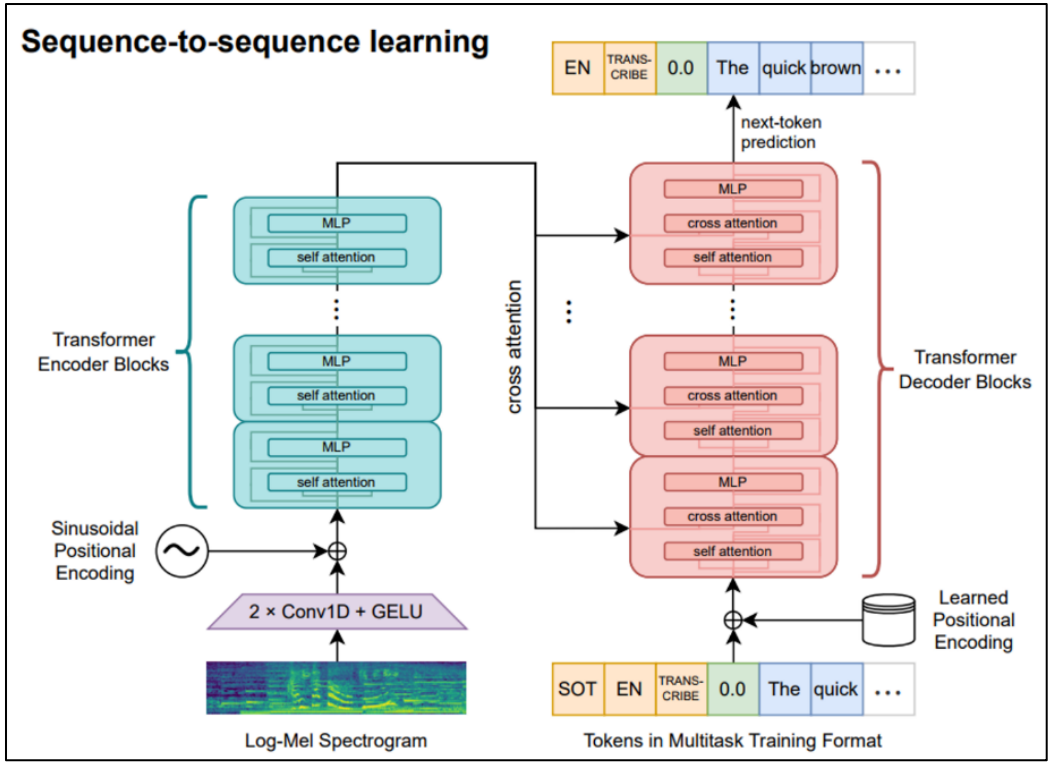
  입력 문장에 대한 정보를 보존하면서

  문장의 의미와 문맥을 파악하여 벡터 형태로 변환

- Decoder :

  인코더의 출력을 기반으로

  인간이 이해할 수 있는 최종적인 형태의 자연어를 출력



3. 프로젝트 전체 구조 및 주요 내용

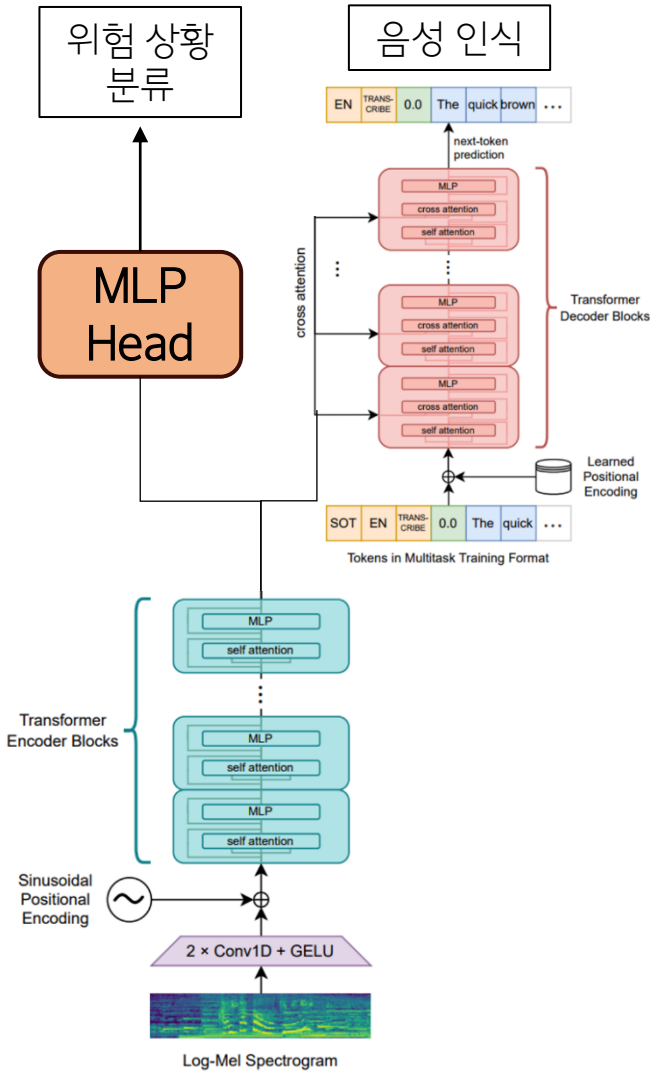
1. Whisper 모델 변형

Whisper 모델의 Encoder 뒤에

위험 상황을 분류하는 MLP Head를 추가

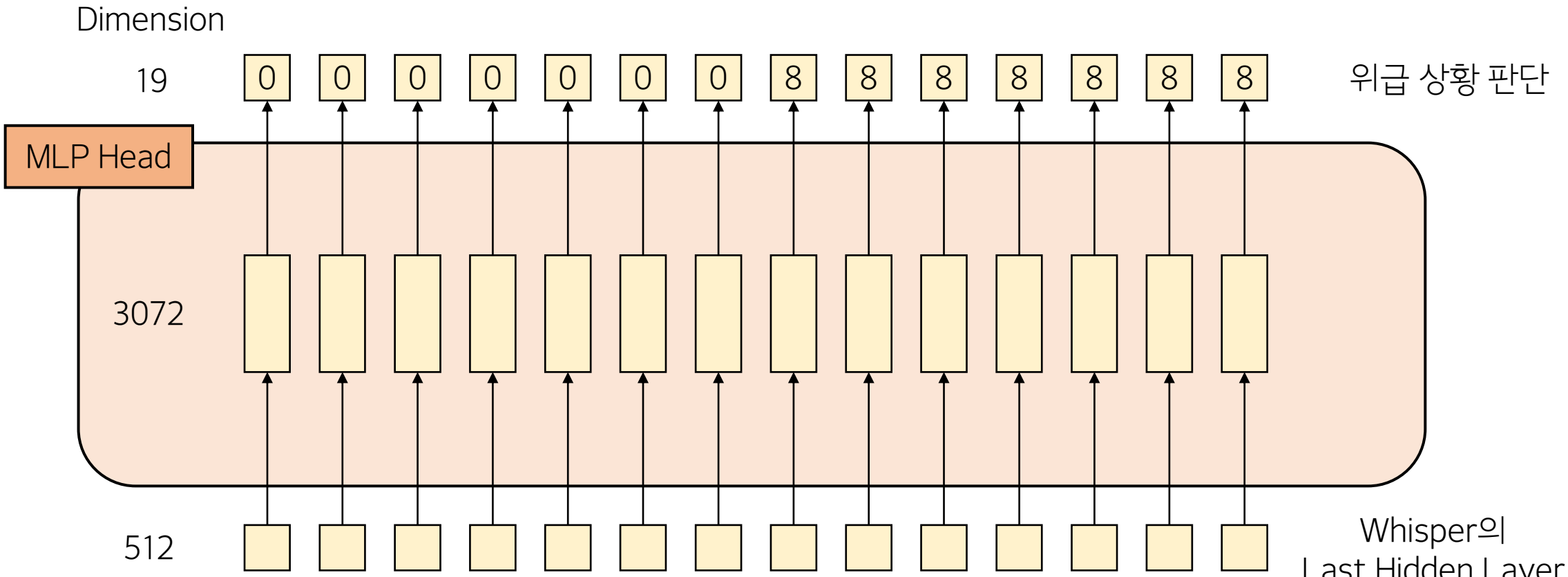
→ 음성 인식과 함께 위급 상황을 분류하는 기능을 수행

[EX] 2. 강도범죄 | “으악 도둑이야!”



3. 프로젝트 전체 구조 및 주요 내용

1. Whisper 모델 변형 : MLP Head



3. 프로젝트 전체 구조 및 주요 내용

1. Whisper 모델 : Text 데이터 처리

사전에 데이터셋을 통해 855개의 위급 상황 리스트를 만들고

위급 상황에 해당하는 text 존재 여부에 따라 위급 상황 판단

기존 데이터는 띄어쓰기, 맞춤법이 올바르게 되어 있지 않아서

Whisper 모델을 돌리면서 어려움이 있었지만,

모든 데이터를 전수 조사하여 틀린 맞춤법을 모두 고침

```
# 맞춤법 교정 사전
correction_dict = {
    '도둑잡아': '도둑 잡아',
    '갇혔나봐': '갇혔나 봐',
    '짜증나게하지마': '짜증나게 하지 마',
    '내몸만지지마': '내 몸 만지지 마',
    '내몸에손대지마': '내 몸에 손대지 마',
    '짜증나게할래': '짜증나게 할래',
    '만지지마세요': '만지지 마세요',
    '짜증나네': '짜증 나네',
    '만지지마': '만지지 마',
    '가스터진거같아': '가스 터진 거 같아',
    '가스터진다': '가스 터진다',
    '어딜손대': '어딜 손대',
    '때리지마세요': '때리지 마세요',
    '어딜만져': '어딜 만져',
    '죽고싶어': '죽고 싶어',
    '손대지마': '손대지 마',
    '갇힌것같아': '갇힌 것 같아',
    '때리지마': '때리지 마',
    '맞고싶냐': '맞고 싶냐',
    '구조해주세요': '구조해 주세요',
    '죽고싶냐': '죽고 싶냐',
    '저 놈 잡아라': '저놈 잡아라',
    '바람이너무세다': '바람이 너무 세다',
```

3. 프로젝트 전체 구조 및 주요 내용

2. CNN 모델 구축

CNN 모델을 통해서

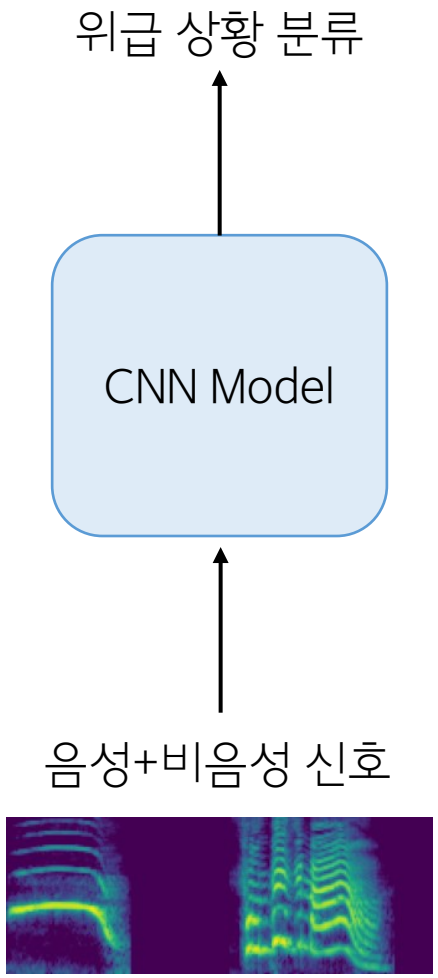
Log-Mel Spectrogram에 대한

위급 상황 분류를 수행

→ Whisper 모델에서

음성 신호를

주로 활용하는 것을 보완

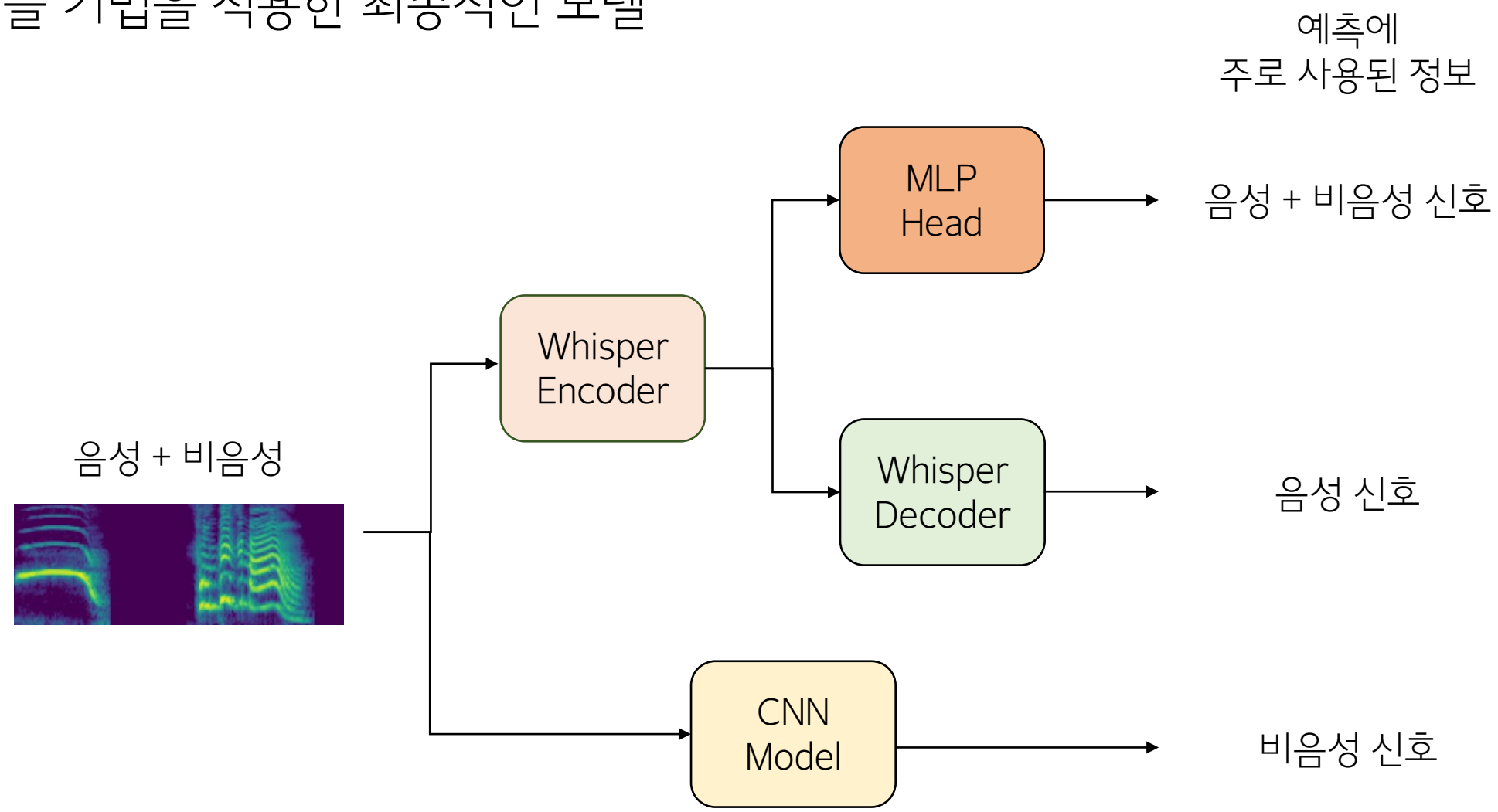


Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 8, 80, 1500]	80
ReLU-2	[-1, 8, 80, 1500]	0
MaxPool2d-3	[-1, 8, 40, 750]	0
Conv2d-4	[-1, 16, 40, 376]	1,168
ReLU-5	[-1, 16, 40, 376]	0
MaxPool2d-6	[-1, 16, 20, 188]	0
Conv2d-7	[-1, 32, 20, 94]	4,640
ReLU-8	[-1, 32, 20, 94]	0
MaxPool2d-9	[-1, 32, 10, 47]	0
Conv2d-10	[-1, 64, 10, 25]	18,496
ReLU-11	[-1, 64, 10, 25]	0
MaxPool2d-12	[-1, 64, 5, 12]	0
Conv2d-13	[-1, 128, 5, 6]	73,856
ReLU-14	[-1, 128, 5, 6]	0
MaxPool2d-15	[-1, 128, 5, 3]	0
Linear-16	[-1, 128]	245,888
Linear-17	[-1, 128]	245,888
ReLU-18	[-1, 128]	0
Dropout-19	[-1, 128]	0
Linear-20	[-1, 64]	8,256
Linear-21	[-1, 64]	8,256
ReLU-22	[-1, 64]	0
Dropout-23	[-1, 64]	0
Linear-24	[-1, 19]	1,235

Total params: 607,763
Trainable params: 607,763
Non-trainable params: 0

3. 프로젝트 전체 구조 및 주요 내용

3. 앙상블 기법을 적용한 최종적인 모델

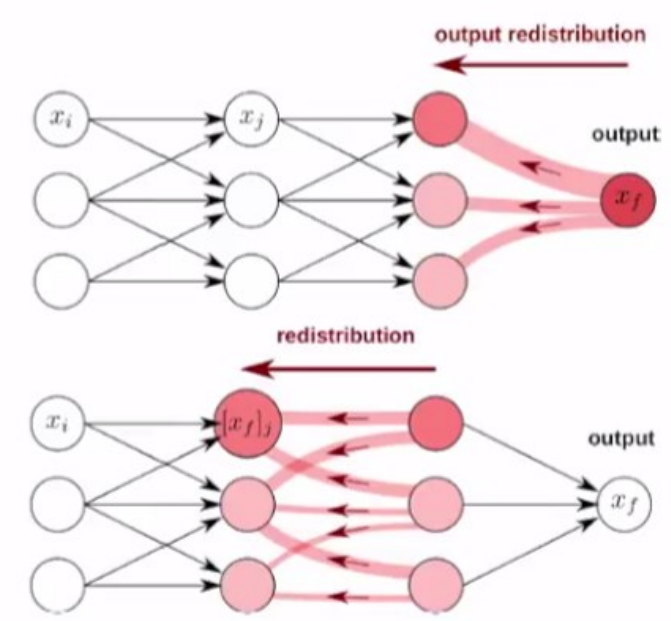


3. 프로젝트 전체 구조 및 주요 내용

4. LRP (Layer-wise Relevance Propagation)

타당성 전파(Relevance Propagation)와 분해(Decomposition) 방법을 사용하여 모델을 해부

출력부터 시작해 타당성 점수(기여도, relevance score)를 입력단 방향으로 계산해 나가며 그 비중을 분배



3. 프로젝트 전체 구조 및 주요 내용

4. LRP 모델 : 음성 분야에서의 LRP

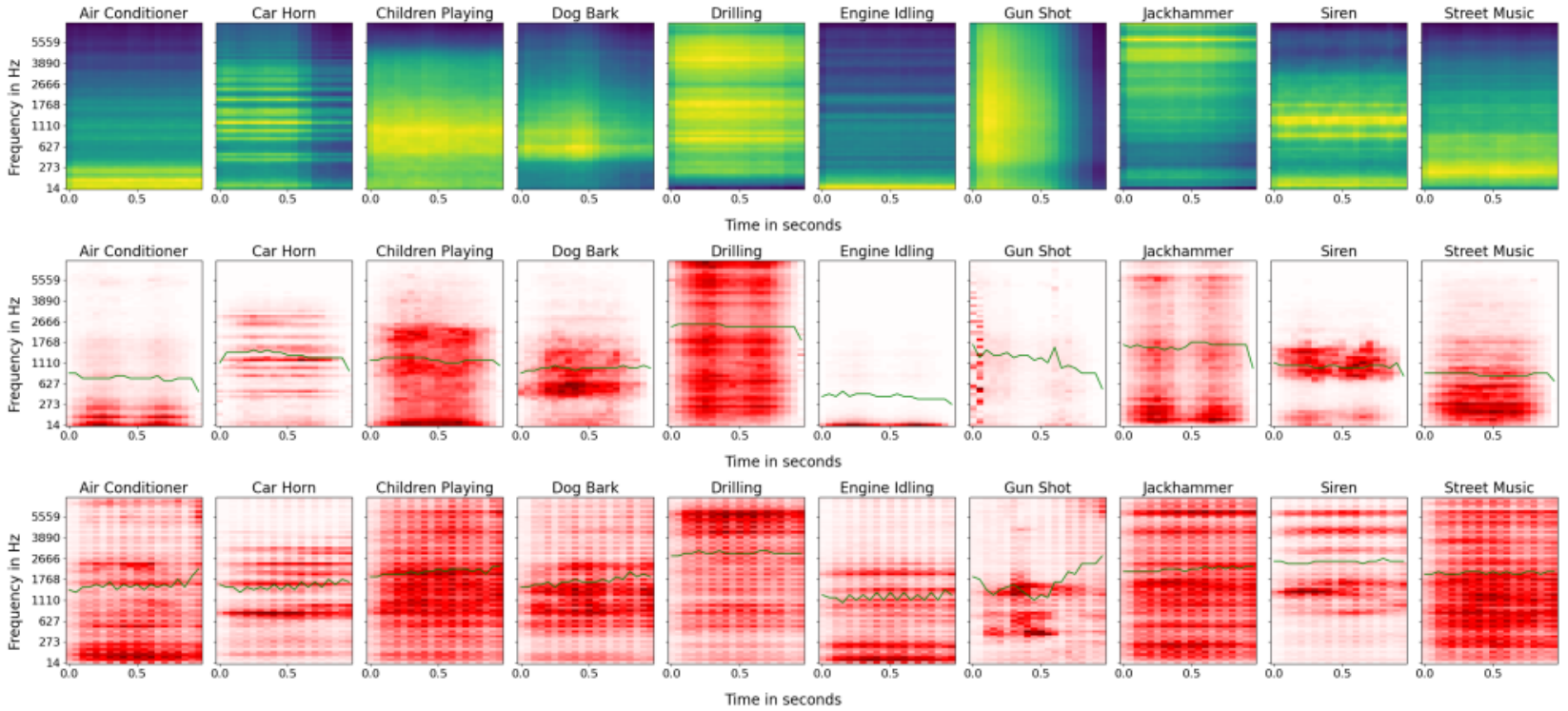
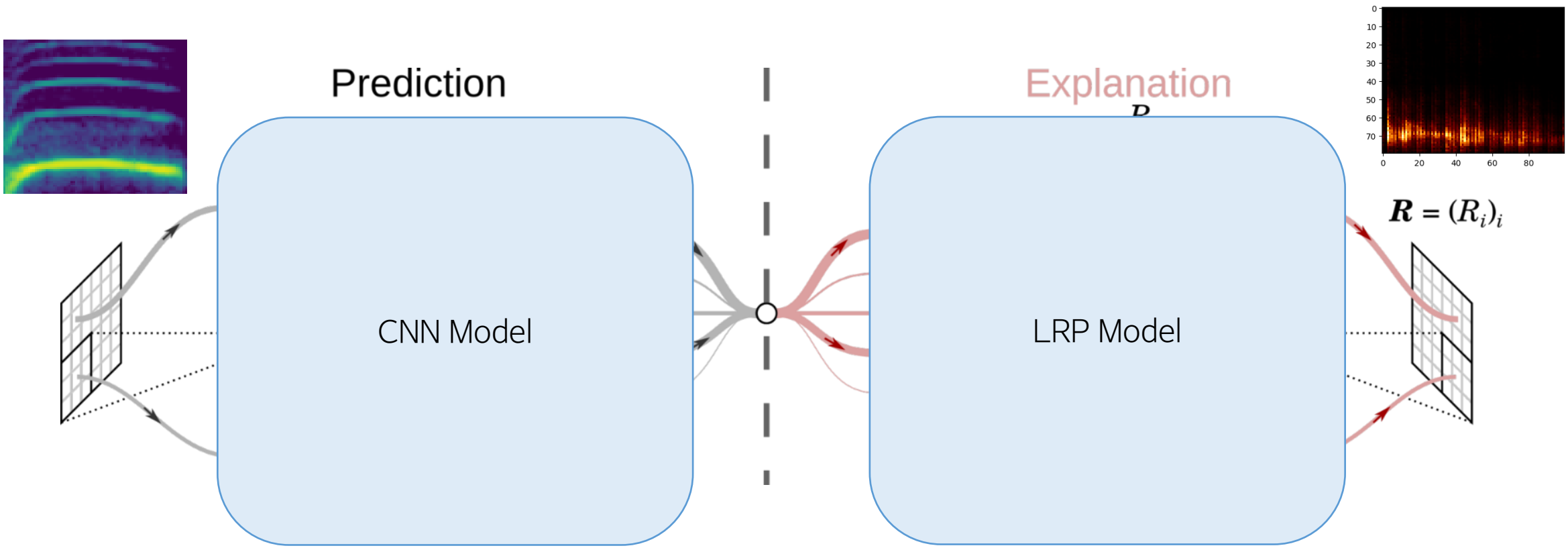


Figure 4: First row: Average spectrograms. Second row: per-class average test set relevance heatmaps for 1DCNN. Third row: per-class average test set relevance heatmaps for YAMNet.

이미지 출처 :XAI-based Comparison of Input Representations for Audio Event Classification, Annika Frommholz, Fabian Seipel, Sebastian Lapuschkin, Wojciech Samek, Johanna Vielhaben

3. 프로젝트 전체 구조 및 주요 내용

4. LRP 모델 : 적용 방식

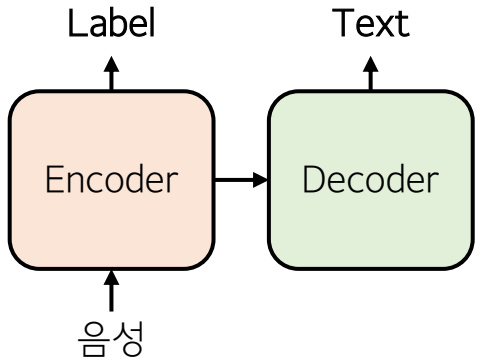


이미지 출처 : https://www.hhi.fraunhofer.de/en/departments/ai/technologies-and-solutions/layer-wise-relevance-propagation.html

4. 프로젝트 모델의 정량적 평가

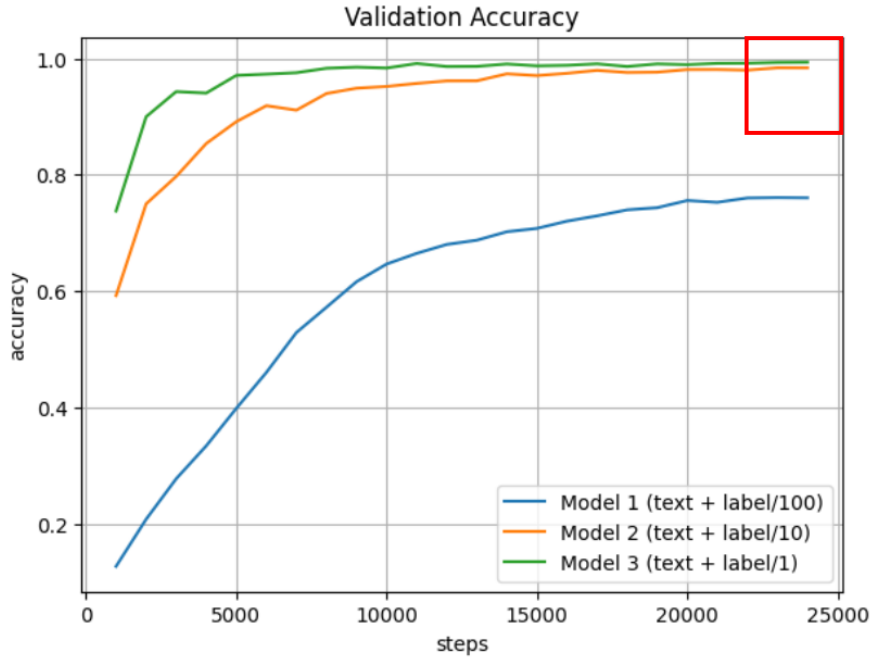
1. Whisper Model (Encoder, Decoder)

- **실험 방법**: Text Loss와 Label Loss 비율을 조절하여 가장 적합한 모델 선정

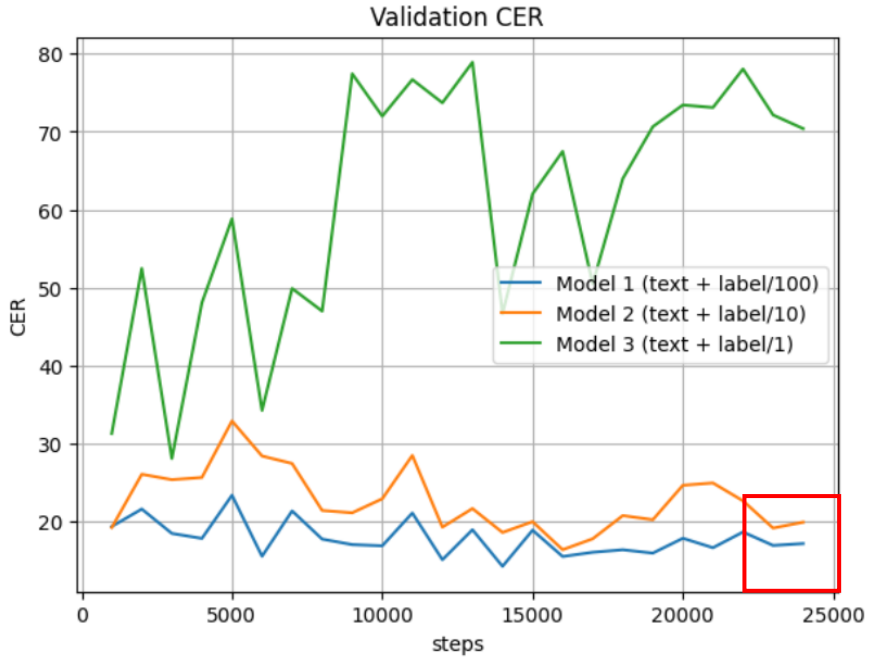


- **실험 결과**:

Accuracy는 높을수록 **Label을 잘 분류함**



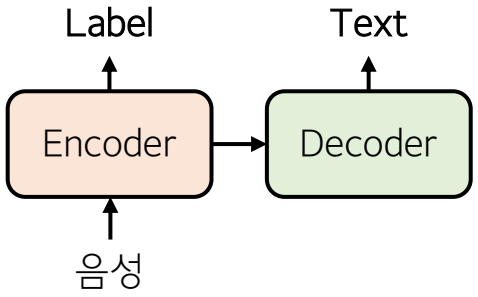
CER은 낮을수록 **Text에 오류가 없음**



Model 2가 가장 적절!

4. 프로젝트 모델의 정량적 평가

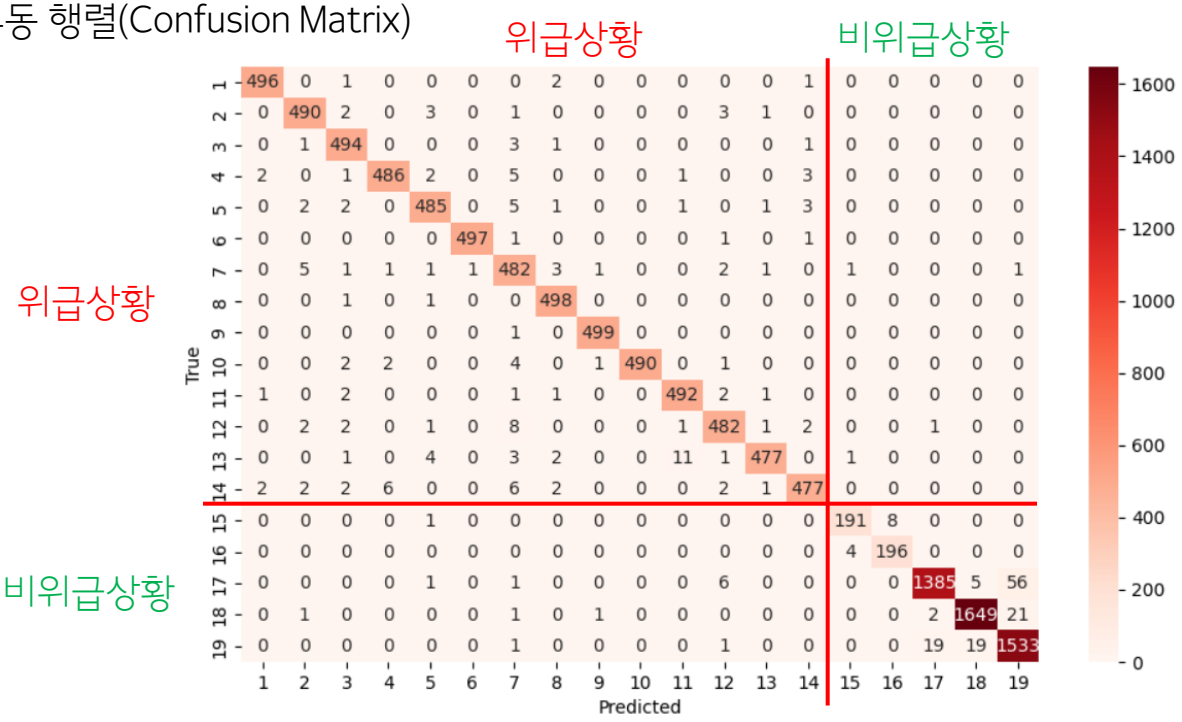
1. Whisper Model (Encoder, Decoder)



- Model 2의 성능

(1) Encoder: 19가지의 위급상황을 97.4%의 정확도로 분류! (위급/비위급상황 이진 분류는 99.85% 정확도로 분류!)

- 혼동 행렬(Confusion Matrix)



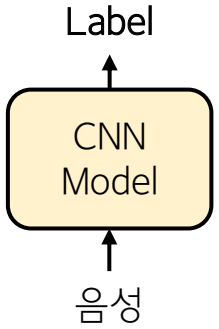
(2) Decoder

- CER을 12.63까지 낮춤
- 텍스트를 통한 위급 상황 이진 분류는 96.21% 정확도로 분류!

```
원본: 내 몸에 손대지 마      | 예측: 내 몸에 손대지 마
원본: 만지지 마              | 예측: 만지지 마만만
원본: 내 몸에 손대지 마      | 예측: 내 몸에 손대지 마
원본: 어딜 손대              | 예측: 어딜 손대
원본: 만지지 마              | 예측: 만지지 마
원본: 손대지 마              | 예측: 손대지 마
원본: 내 몸 만지지 마        | 예측: 내 몸 만지지 마
```

4. 프로젝트 모델의 정량적 평가

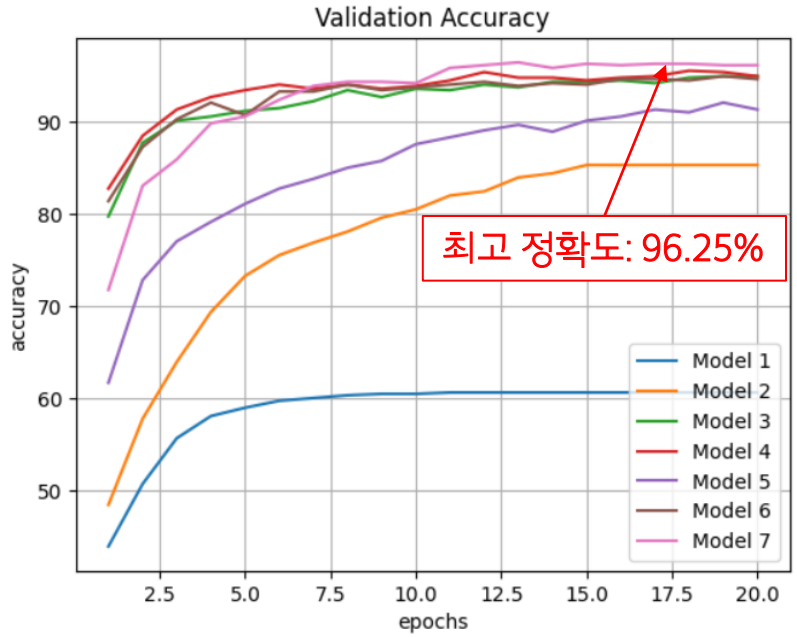
2. CNN Model



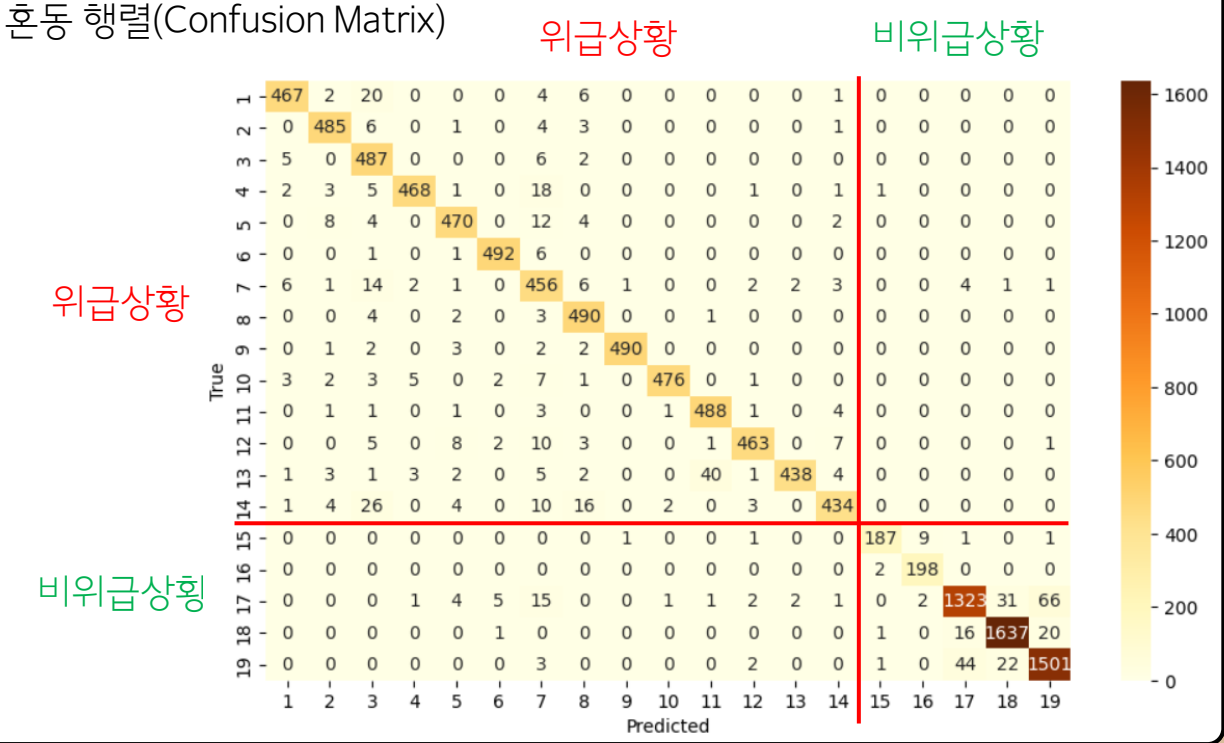
- **실험 방법**: 모델의 구조(Architecture)와 하이퍼파라미터(Hyper-parameters)를 조절하여 가장 정확도가 높은 모델 선정

- **실험 결과**: 7개 이상의 모델 결과, 가장 높은 정확도를 보인 Model 7 선택,

  Test 정확도 94.61%, 이진 분류 정확도 99.60%



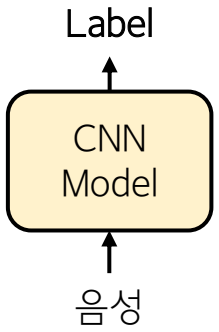
- 혼동 행렬(Confusion Matrix)

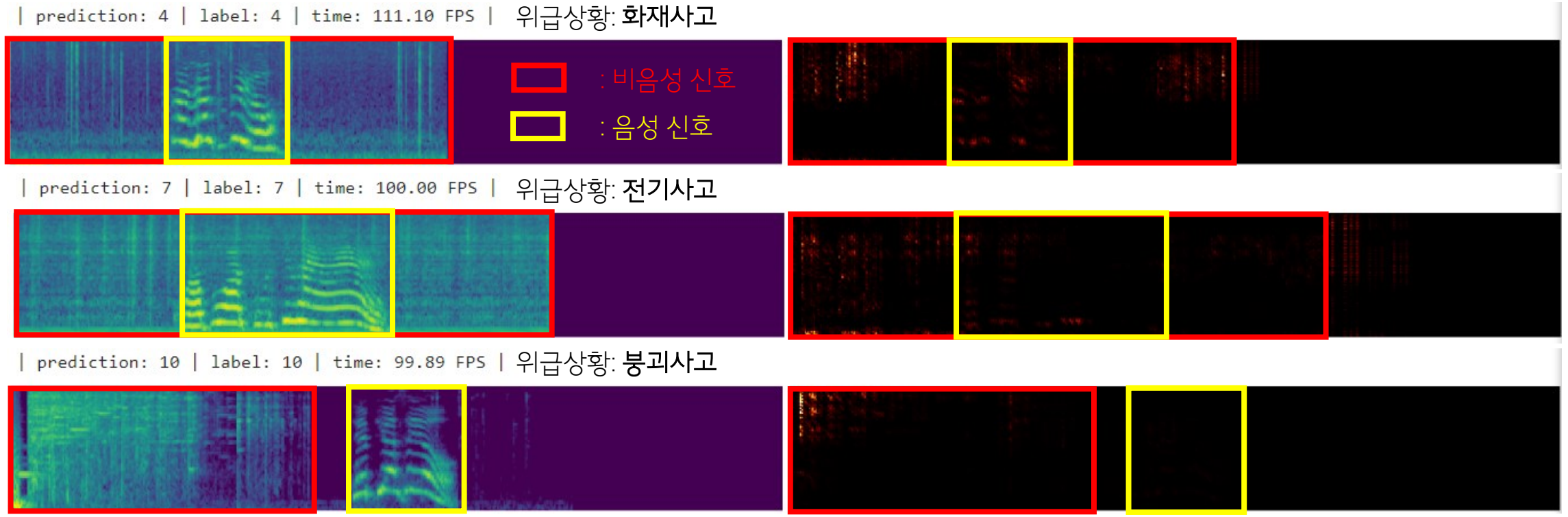


4. 프로젝트 모델의 정량적 평가

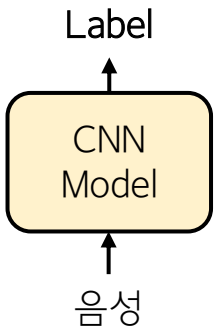
2. CNN Model

- 실제로 CNN Model은 음성 신호보다 비음성 신호를 고려하여 분석하는가?
- **LRP**(Layer-wise Relevance Propagation)를 이용한 분석:





4. 프로젝트 모델의 정량적 평가

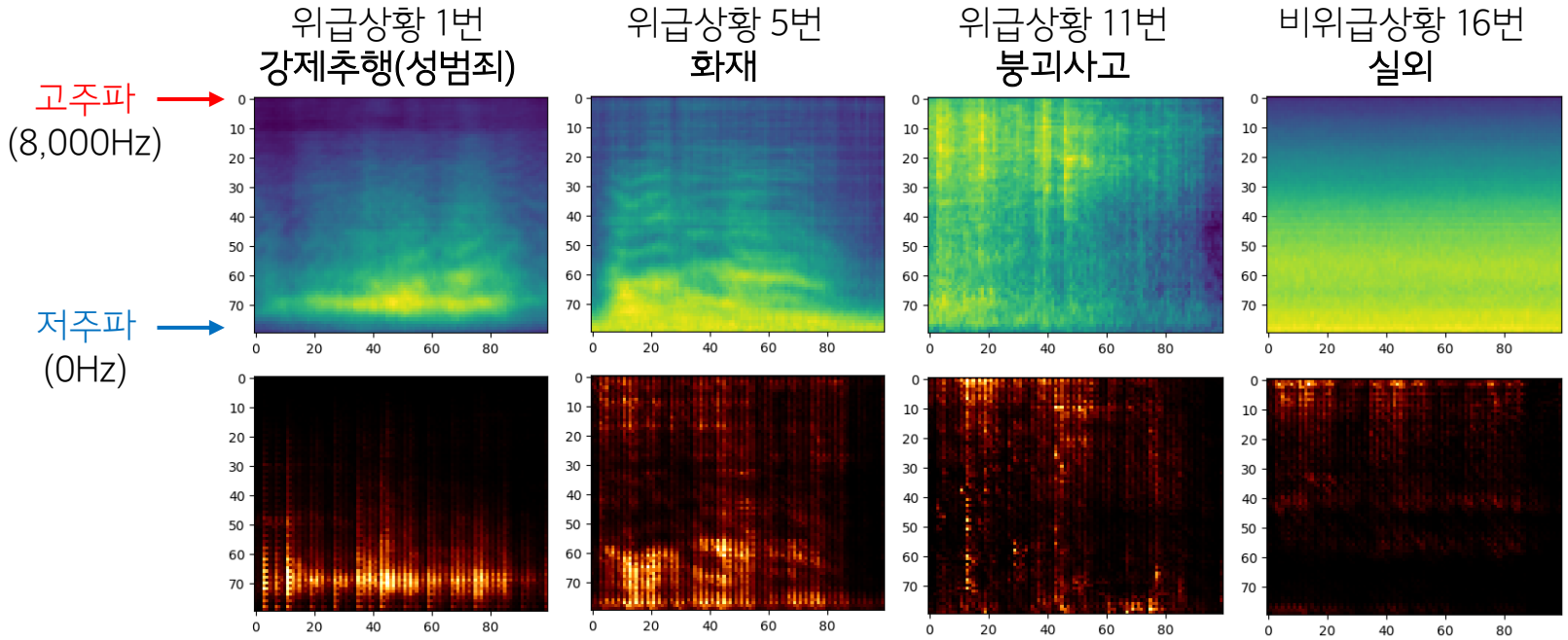


2. CNN Model

- **LRP**(Layer-wise Relevance Propagation)를 이용한 분석:

- CNN Model이 각 위급/비위급상황에서 어느 주파수 범위를 위주로 살펴보았는지 알 수 있음.

- 각 상황의 **멜-스펙트로그램(Mel-spectrogram)**과 **LRP 결과**의 **평균** (약 100장의 평균)



4. 프로젝트 모델의 정량적 평가

3. Ensemble Model

- 1) CNN Model, 2) Whisper Encoder, 3) Whisper Decoder를 결합한 **앙상블 모델**
- 3개의 모델 중 **2개의 모델**이 위급상황이라고 판단할 경우 → 위급상황으로 분류

→ Ensemble Model은

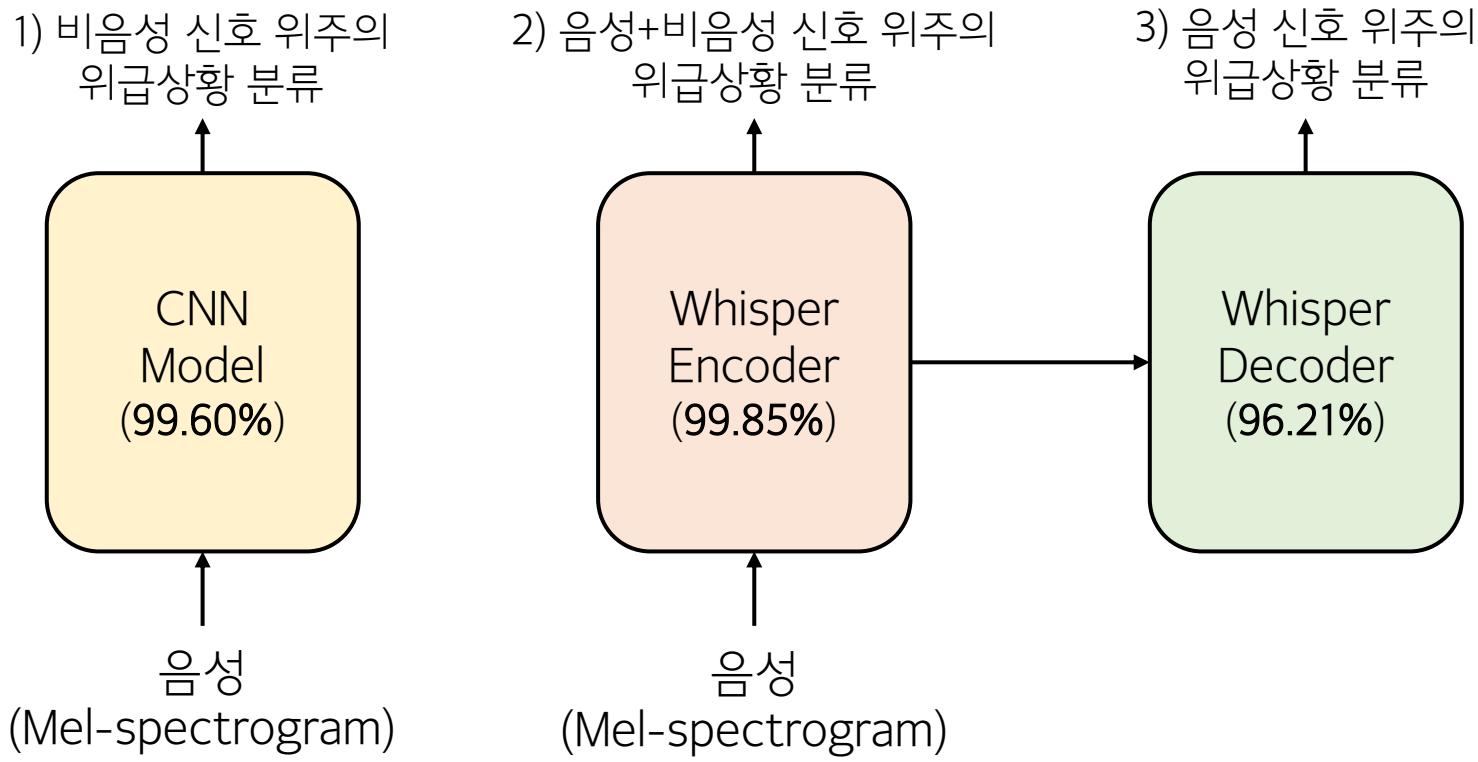
최종적으로 위급/비위급상황 이진분류에서

99.93%의 정확도를 갖는다!

```
acc = 0
n = len(LABELS_LIST)
for label, cnn, whisper_enc, whisper_dec in zip(LABELS_LIST, CNN_LIST,
                                                WHISPER_ENCODER_LIST,
                                                WHISPER_DECODER_LIST):
    pred = (sum([cnn, whisper_enc, whisper_dec]) >= 2)
    if label == pred:
        acc += 1 / n

print(f"Final Accuracy: {acc*100}%")
```

```
Final Accuracy: 99.93389522390196%
```



4. 프로젝트 모델의 정량적 평가

4. Demo

		정답	CNN Model	Whisper Encoder	Whisper Decoder
1		위급상황: 10. 낙상 *(물건이 떨어지는 소리) (약한 신음 소리)*	위급상황 : 14. 도움요청	위급상황 : 10. 낙상	비위급상황 *악*
2		비위급상황: 15. 실내 *(잔잔한 음악) 어 여보세요?*	비위급상황 : 15. 실내	비위급상황 : 15. 실내	비위급상황 *너 여보세요*
3		위급상황: 14. 도움요청 *살려주세요!*	위급상황 : 4. 폭력범죄	위급상황 : 14. 도움요청	위급상황 *살려주세요*
4		비위급상황: 17~19. 대화 소음 *이 모델이 '살려주세요'라고 얘기하는 것만으로 위급상황이라고 판단할까?*	비위급상황 : 17. 대화 소음	비위급상황 : 19. 대화 소음	위급상황 *이 모델이 살려주세요라고 얘기하는 것만으로 육급 상황이라고 판단할까?*

5. 프로젝트의 한계점 및 보완 과제

1. 한계점

(1) 데이터의 한계점

- AI Hub의 위급상황 데이터는 실제 위험상황이 아닌 **인위적인 위험상황**이기 때문에 실생활 적용의 어려움이 있음.
- **제한된 컴퓨팅 자원**으로 인해 500GB의 데이터를 모두 활용하지 못함.

(2) 시간적 한계

- CNN 모델, Whisper 모델을 한 번 학습시키는 데에 2~6시간
- 모델 실험 시간이 오래 걸려서 많은 실험을 해볼 수 없었음.

(3) 모델의 결과 종합 방식

- CNN 모델과 Whisper 모델의 위험상황 분류 결과가 다를 때 둘의 결과를 종합하는 방식을 개선할 필요가 있음.
- 위급상황을 비위급상황이라고 판단할 가능성이 높아짐.

5. 프로젝트의 한계점 및 보완 과제

2. 보완 과제

(1) 최적의 모델 구조와 하이퍼파라미터 탐색

- Whisper Decoder의 정확도(96.21%)가 다른 모델의 정확도(99.60%, 99.85%)에 비해 낮은 편
- MLP Head 구조 변경 및 Loss 비율의 조절로 보완이 가능할 것으로 예상됨.

(2) 시각별로 위급상황 판단할 수 있도록 데이터 보강

- 원래는 시각별로 위급상황이 발생한 위치를 탐색하려고 했음.
- 그러나 컴퓨팅 자원의 한계 및 위급상황 구간 파악의 어려움으로, 데이터를 추가하지 못함.
- 추후에 시각별로 위급상황을 판단하도록 개선할 예정

(3) Transformer XAI를 통한 모델의 설명력 증가

- Whisper Encoder에서 음성 신호 위주로 추론했음을 보이기 위해 사용 가능