



Real-Time Scream Detection

using Deep Learning



Eye-SCREAM

신현경 신훈철 유기운 임채빈 최세영



CONTENTS

- Idea
- Data
- Model
- Demo
- Service
- Summary



Idea



Idea

비명소리를 감지하여
위험한 상황을 빠르게 대처하기 위한
프로젝트



Idea

'방관자 효과는 없었다' 비명 듣고 달려가 성폭행 위기 여성 구한 행인 4명

[사회현장] 버닝썬 'VIP룸' 폭로..."여자 비명소리 들려도 무시"

지하철5호선 女화장실에 '비명' 감지장치...긴급상황 실시간 전달

지하철 성범죄 주요 발생 장소지만 사생활 보호를 위해 폐쇄회로텔레비전(CCTV)을 설치할 수 없었던 여자화장실에 신형 장치가 설치된다.

이 장치(세이프 메이트·Safe Mate)는 비명을 감지해 실시간으로 긴급 상황을 알려준다. 비명이 장치에 감지되면 화장실 입구 경광등이 울리고 역 직원 휴대전화로 상황이 전달된다. 이 상황을 경찰에 전송하는 장치도 구축될 예정이다.

출처 - 네이버뉴스



Idea

- 소리 데이터에 대한 관심
 - 특징적인 비명소리에 대한 탐구
 - 시각적 효과 만큼이나 청각적 효과 또한 강력함
 - 장애물을 통과하여 전달되는 특징

- 긴급상황 실시간 전달
 - 긴급상황 대처 방안으로 유력

- 위험한 상황에서의 비명소리 빈번히 발생
 - 보호받지 못하는 사각지대에서 발생하는 비명소리 또한 감지

<표 3-15> 실현가능한 방법으로 선택한 내용

구분	건수	비율
경찰신고, 인력배치, 신고처 설치, CCTV설치	65	30.0%
소리지르기, 직접이야기, 주변사람에게 도움, 버스기사에게 도움	60	27.8%
자리어둥, 즉시 하차	41	19.0%
경보벨 설치, 휴대폰 인증, 호신술 이용	35	16.2%
성범죄예방 홍보 등	6	2.8%
대중교통혼잡완화	5	2.3%
여성전용칸	4	1.9%
합계	216	100.0%

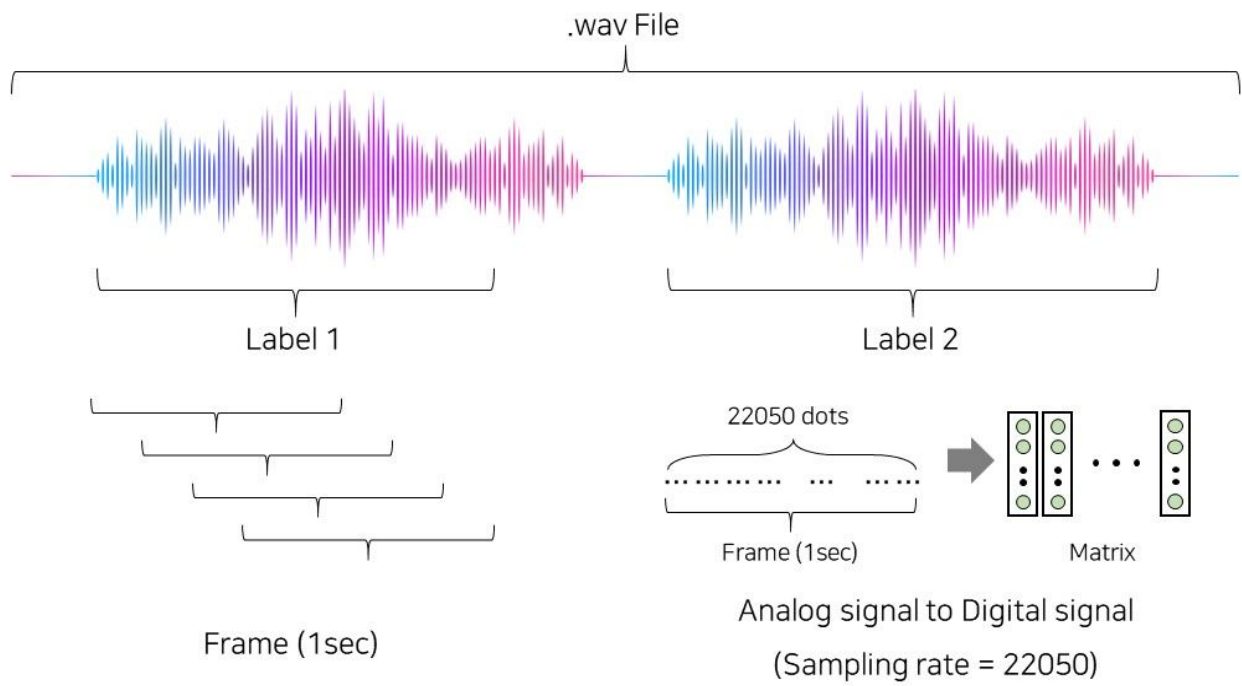


Data

- Overview
- Gathering
- Labeling
- Augmentation
- Preprocessing
- Final Feature



Data Overview






Data Gathering



- Youtube Data
 - 'Screaming' 단어가 들어가는 검색어 키워드 중심으로 사이트 주소 Crawling
 - 23개 키워드에 대해 총 2078개 동영상 사이트 주소를 수집
 - 동영상 수집 후 음성파일 추출
 - 비명 소리만 있지 않고 다양한 소리도 섞여 있어서 같이 사용
- Google Search Data
 - '효과음' 키워드 중심으로 검색해 다양한 블로그와 카페에서 무료로 제공된 데이터 수집
 - 비명소리 뿐만 아니라 비명소리와 헛갈릴 수 있는 소방차 사이렌, 경찰차 사이렌 등 비슷한 기계음과 환호소리, 토하는 소리, 트림 소리 등 비슷한 사람소리도 같이 구분하기 위해 수집
- Self-recorded Data
 - 직접 비명을 녹음하여 수집

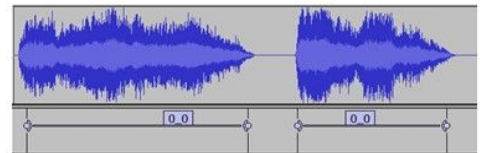
Data Labeling

• Labeling

- .wav 소리 데이터를 직접 듣고 시작과 끝 비명여부를 .txt 형태로 저장
- Using Audacity 

• Labeling Rule

- 비명인 경우 1, 비명이 아닌 경우 0 → **Binary classification**
- 0의 경우, classifier의 error analysis를 위해 언더바(_) 사용
- Rule
 - 1: 비명. 롤러코스터의 비명, 욕설이 섞인 비명을 포함한 소리
 - 0_0: 환호, 소리지르기, 웃음 소리, 소프라노 소리 등 들을 때 헛갈릴 만한 사람 소리
 - 0_1: 새소리, 차소리 등 사람이 내지 않는 기계가 헛갈릴 만한 높은 소리
 - 0_2: 낮은 소리를 포함한 나머지 배경소리



datafile19.txt - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말

0.289088 5.014982 0_0

6.070293 9.245245 0_0

< Labeling using Audacity>

Data Labeling

- Efficient-Labeling System

Segmentation with threshold



Predict per frame



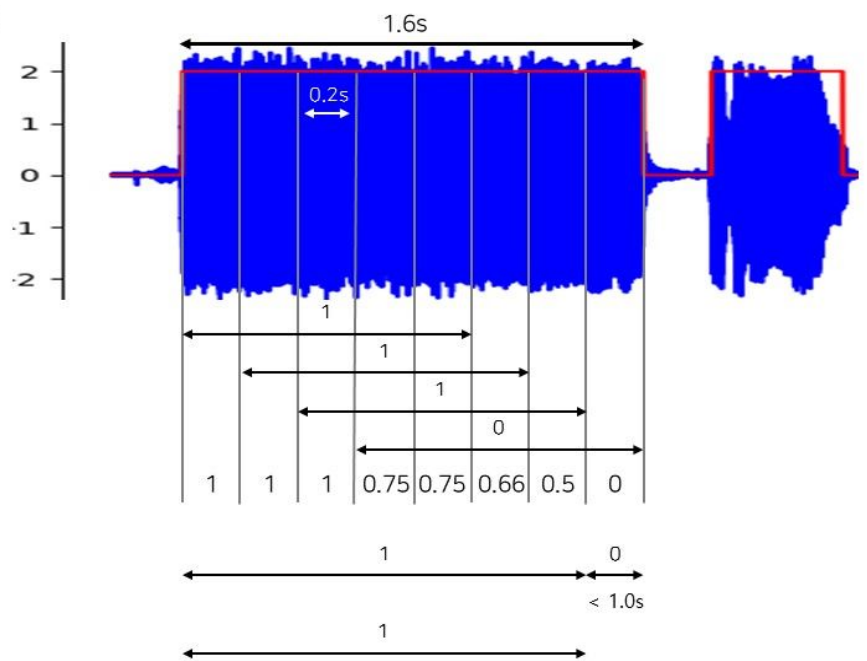
Average per segment



Model-decision



Listen and correct in person



Data Labeling

- Efficient-Labeling System

Segmentation with threshold



Predict per frame



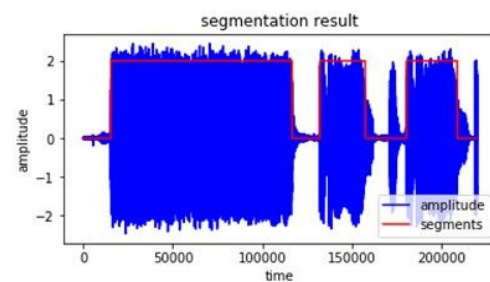
Average per segment



Model-decision



Listen and correct in person



```
[[0.69904762 5.19904762 1. ]  
 [5.98526077 6.98526077 0. ]  
 [8.18054422 9.18054422 1. ]]
```

▶ 0:04 / 0:04 ————— 🔊 ⋮

predicted label : 1
1

▶ 0:01 / 0:01 ————— 🔊 ⋮

predicted label : 0
1

▶ 0:01 / 0:01 ————— 🔊 ⋮

predicted label : 1
1
datafile13.txt saved



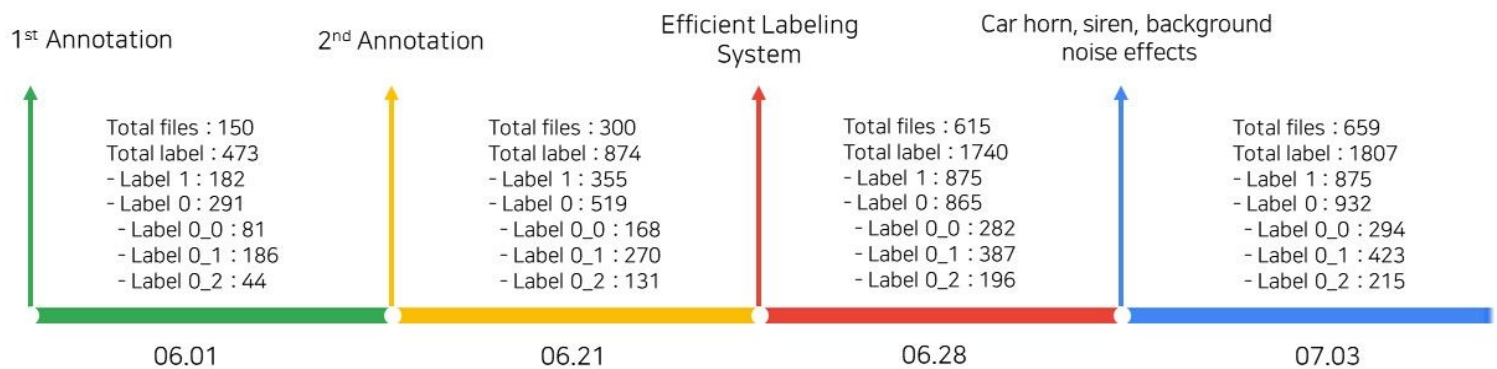
Data Labeling

- Dataset Timeline

- Total files : 659
 - Youtube audio data : 500
 - Google search data : 159

- Total labeled segment : 1807

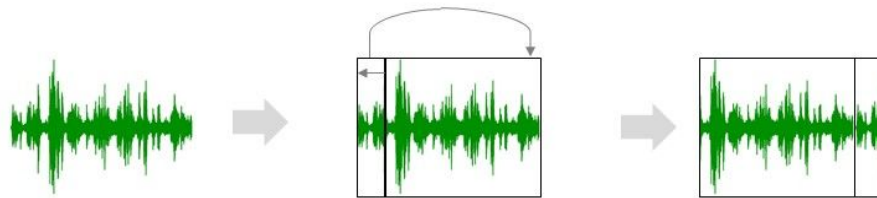
- Label 1 : 875
- Label 0 : 932
 - Label 0_0 : 294
 - Label 0_1 : 423
 - Label 0_2 : 215



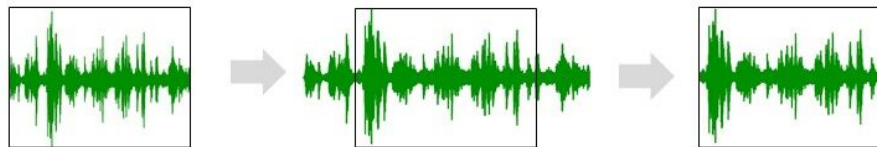


Data Augmentation

- Time Shifting



- Speed Tuning



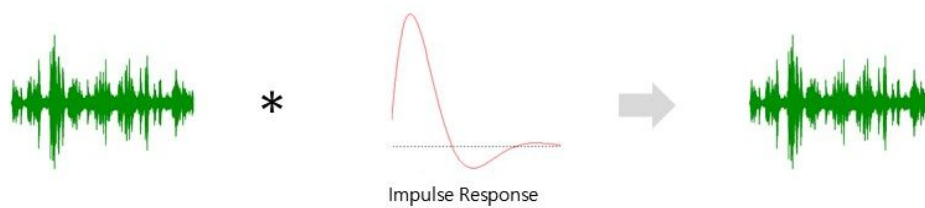
- Mix background noise





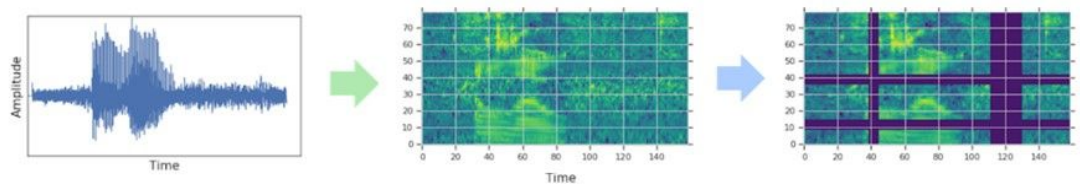
Data Augmentation

- Convolution with impulse response



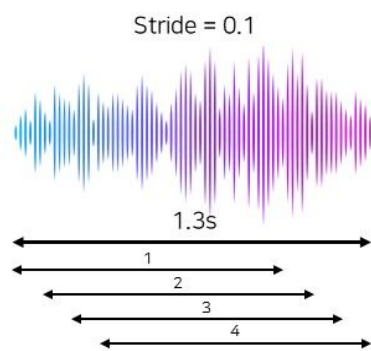
- SpecAugment

1. Time warping
2. Frequency masking
3. Time masking

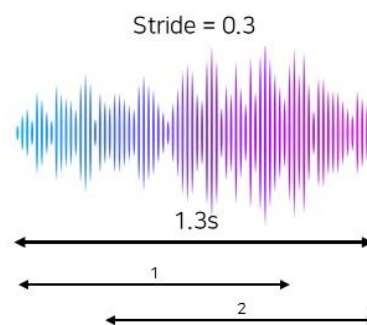


Data Preprocessing

- Frame preprocessing
 - Wav file → Frame
 - 1 Frame : 1 second (22050 Hz)
 - Various frame stride parameter
 - 0.1
 - 0.2
 - 0.3



Sampled Frame





Data Preprocessing

- Feature Extraction
 - Mel Spectrogram
 - Compute mel-scaled spectrogram
 - We chose 64-mels spectrogram
 - Log – Mel Spectrogram
 - Compute log-mel-scaled spectrogram
 - Log has the effect of perceiving like human`s ear
 - We chose 64-mels spectrogram



Data Preprocessing

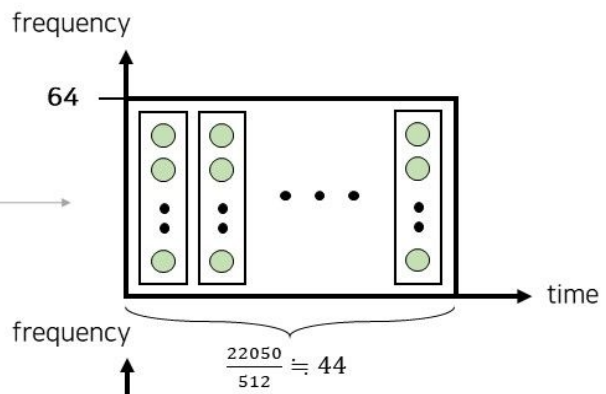
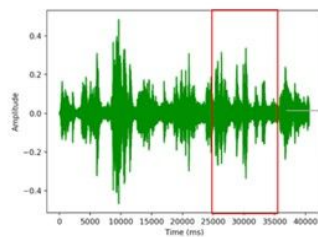
- Feature Extraction
 - Features
 - Mel
 - Log Mel
 - Various values of n_fft, hop size parameter
 - 512
 - 2048
 - 4096
 - Feature shape

hop_size	512	2048	4096
Mel	(N, 64, 44)	(N, 64, 11)	(N, 64, 6)
log Mel	(N, 64, 44)	(N, 64, 11)	(N, 64, 6)

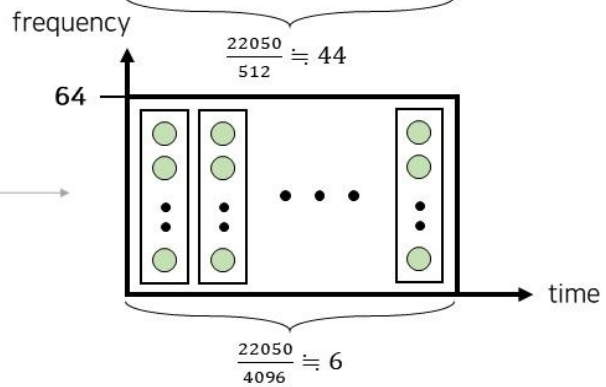
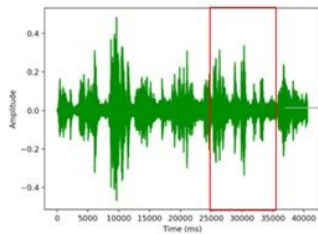
Data Preprocessing

- Feature Extraction

- Hop size, $n_fft - 512$

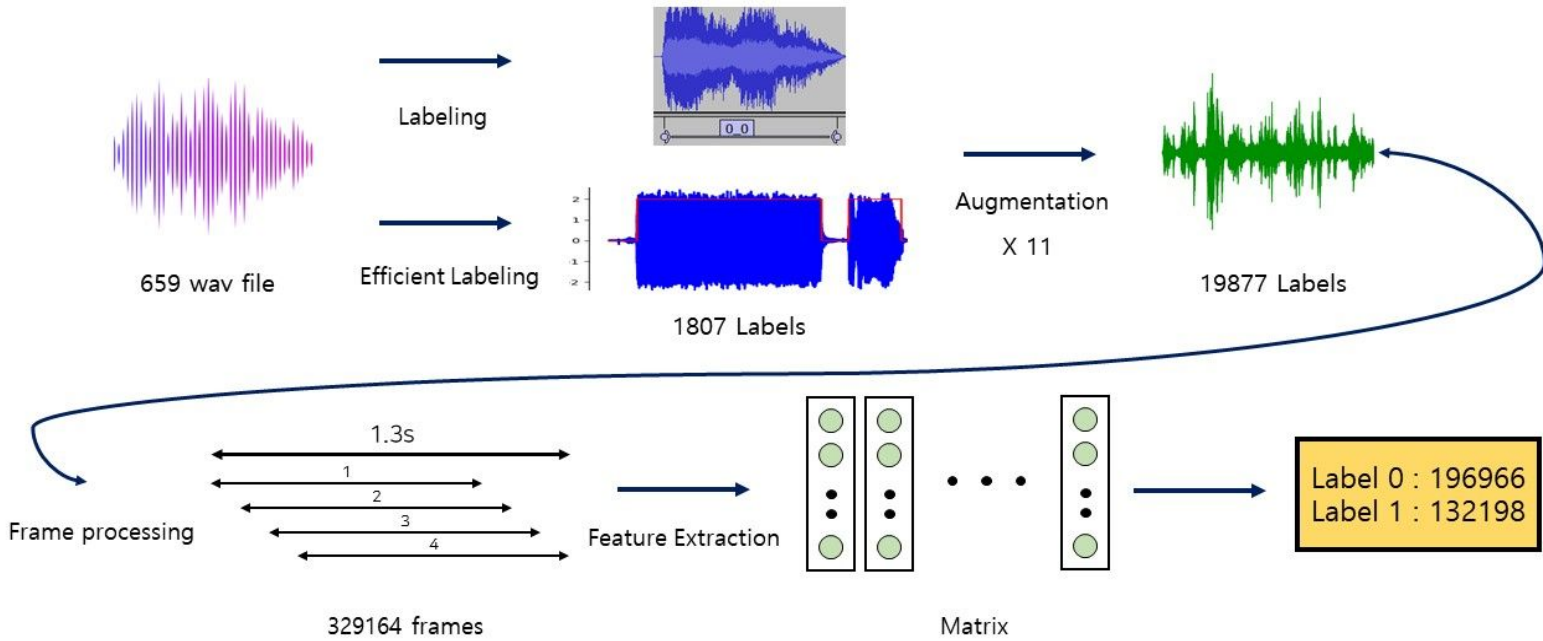


- Hop size, $n_fft - 4096$





Final Feature





Model

- Structure
- Experiment
- Model Size Comparison
- Hyperparameter tuning

Model Structure

- Model Types

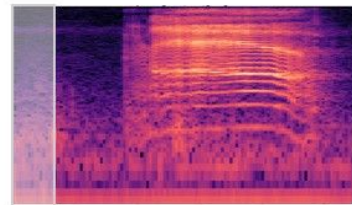
- Time-independent CNN

- Type A : $1 * N$ kernel (Time-based)
- Type B : $N * 1$ kernel (Frequency-based)
- Type C : $M * N$ kernel

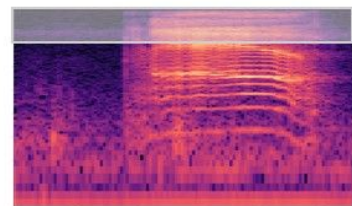
- Time-dependent CNN

- CNN - RNN (LSTM)

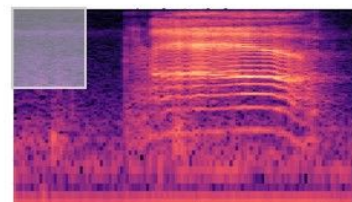
Type A



Type B



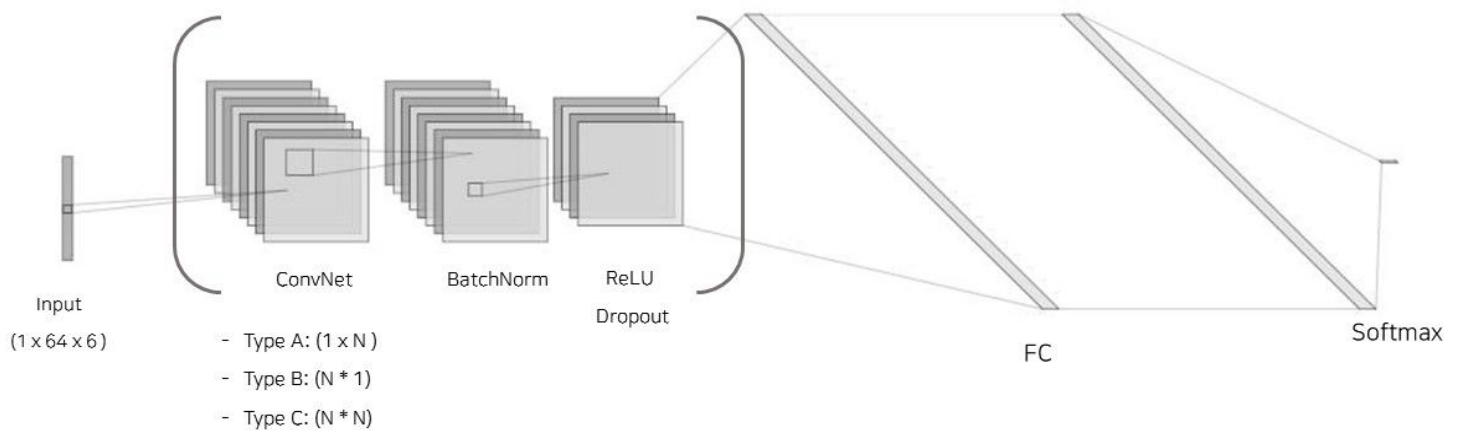
Type C





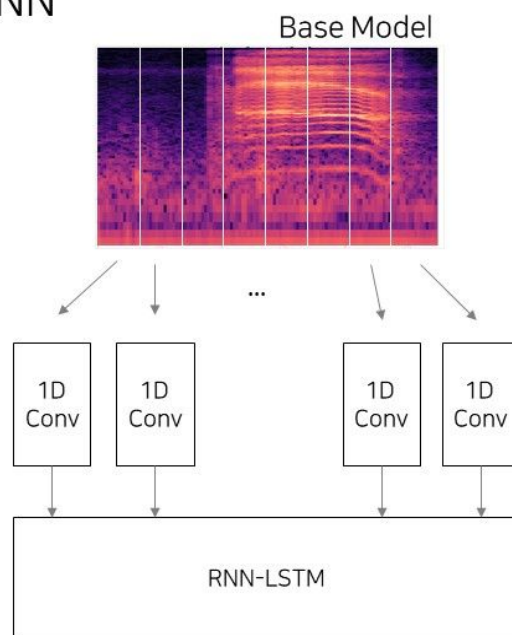
Model Structure

- Time-Independent CNN



Model Structure

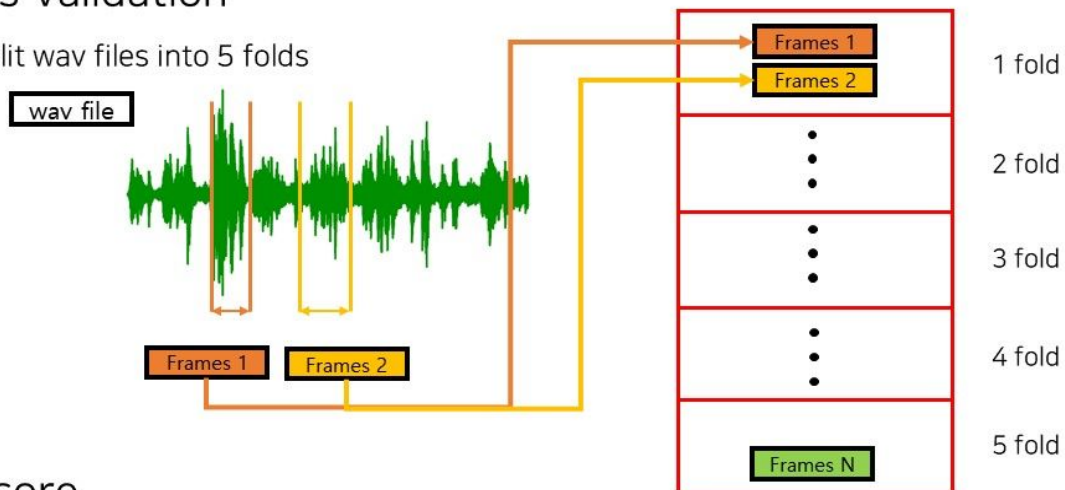
- Time-dependent CNN



Experiment - Metric

- Cross Validation

- Split wav files into 5 folds



- F1 score

- The harmonic mean of precision and recall
- Better metric than accuracy for unbalanced data



Experiment - Feature Pruning

- Feature의 경우의 수가 많다.
- 효율적인 실험을 위해 Feature Pruning 필요
- Input shape에 따른 속도 차이도 컸음

	hop size	512			2048			4096		
	Stride	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3
Mel	Basic	0.654	0.456	0.440	0.658	0.593	0.417	0.647	0.585	0.425
	Aug	0.750	0.744	0.733	0.757	0.738	0.744	0.752	0.754	0.751
	SpecAug	0.761	0.754	0.727	0.778	0.735	0.752	0.767	0.757	0.753
Log-Mel	Basic	0.471	0.100	0.295	0.410	0.170	0.358	0.489	0.327	0.314
	Aug	0.742	0.705	0.733	0.744	0.665	0.68,83	0.731	0.734	0.715
	SpecAug	0.727	0.734	0.559	0.738	0.738	0.63,83	0.736	0.724	0.695



Experiment - Feature Pruning

- Feature의 경우의 수가 많다.
- 효율적인 실험을 위해 Feature Pruning 필요
- Input shape에 따른 속도 차이도 컸음

	hop size	512			2048			4096		
	Stride	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3
Mel	Basic	0.654	0.456	0.440	0.658	0.593	0.417	0.647	0.585	0.425
	Aug	0.750	0.744	0.733	0.757	0.738	0.744	0.752	0.754	0.751
	SpecAug	0.761	0.754	0.727	0.778	0.735	0.752	0.767	0.757	0.753
Log-Mel	Basic	0.471	0.100	0.295	0.410	0.170	0.358	0.489	0.327	0.314
	Aug	0.742	0.705	0.733	0.744	0.665	0.68,83	0.731	0.734	0.715
	SpecAug	0.727	0.734	0.559	0.738	0.738	0.63,83	0.736	0.724	0.695



Experiment - Model Pruning

- ~ 20 Epoch (F1 score as metric)

	0 Fold	1 Fold	2 Fold	3 Fold	4 Fold	Average
<u>Time-independent</u>						
A_Depthwise	0.875	0.87	0.837	0.834	0.864	0.856
A	0.928	0.928	0.825	0.843	0.938	0.892
B	0.884	0.764	0.808	0.810	0.836	0.812
C	0.949	0.886	0.842	0.813	0.948	0.888
<u>Time-dependent</u>						
C-RNN	0.863	0.880	0.821	0.792	0.861	0.843



Model Size Comparison

METRICS	A	A_Depthwise	B	C	A_Deep
FLOPS (actual compute)	1,179K	771K	2,785K	3,865K	4423K
# of Parameters	165K	297K	533K	420K	659K
Size (Mb)	0.908	1.499	2.724	2.212	3.065
Parameter (kB)	5294	9511	17066	13425	21074
Forward/Backward (kB)	2314	3052	5771	5116	4629
Input (bits)	12288	12288	12288	12288	12288




Model Size Comparison

METRICS	A	A_Depthwise	B	C	A_Deep
FLOPS (actual compute)	1,179K	771K	2,785K	3,865K	4423K
# of Parameters	165K	297K	533K	420K	659K
Size (Mb)	0.908	1.499	2.724	2.212	3.065
Parameter (kB)	5294	9511	17066	13425	21074
Forward/Backward (kB)	2314	3052	5771	5116	4629
Input (bits)	12288	12288	12288	12288	12288



Hyperparameter Tuning

- Random Search of 30 epochs using,
 - Learning rate : (1e-5, 5e-2)
 - Dropout rate : (0.0,0.7)
 - Adam parameters:
 - Betas : (0.9, 0.99)
 - Eps : 1e-8
 - L2 Regularizer : (0.0, 1e-6)
- 
- Learning rate: 5.95e-3
 - Dropout rate: 0.48
 - (Betas: 0.914, Eps: 1e-08)
 - Regularizer: 8.86e-7

5-fold average metrics on final dataset:

F1 score: 0.936, Acc. : 0.947, Recall: 0.945, Precision: 0.928

5-fold average metrics on previous dataset (for comparison):

F1 score: 0.901 (-3.5%p), Acc. : 0.917(-3.0%p), Recall: 0.909(-3.8%p), Precision: 0.891 (-3.9%p)



Demo



Demo





Service

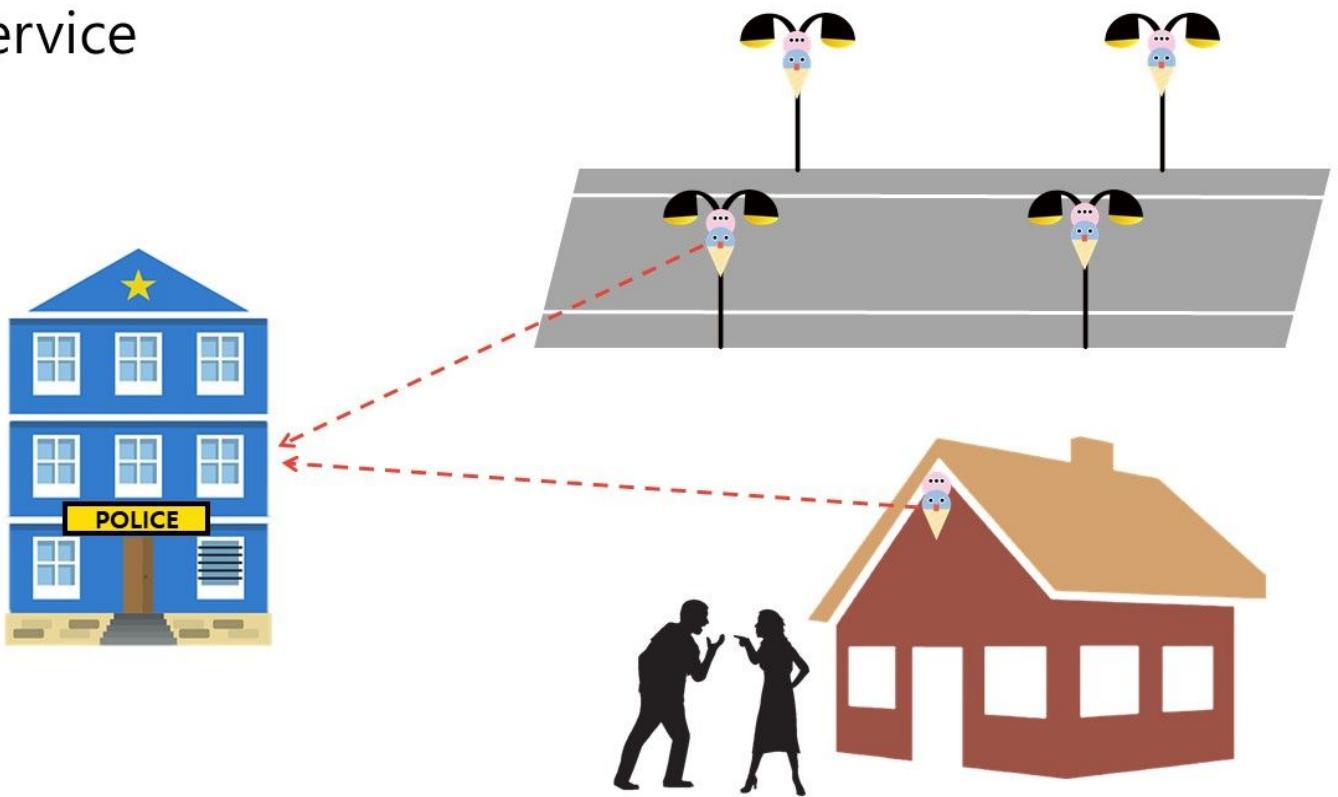


Service

- 골목길 가로등에 설치
 - 범죄율이 높은 지역의 범죄 발생율 감소 효과 기대
 - 범죄 검거를 위한 인력 비용 감소
 - 실시간 서비스 구축
- 가정 집에 설치
 - 자취방에서 일어날 수 있는 범죄 방지
 - 맞벌이 부부가 사는 집의 아이 위험상황 감지
 - 가정 폭력, 아동 학대의 빠른 검거 및 조치 가능



Service





Summary



Conclusion

- 결론 및 의의
 - 소리 데이터를 경험할 수 있는 좋은 기회
 - 실제 서비스화 될 가능성이 존재
 - 프로젝트 내 효율적인 시스템 직접 구축
 - Failure Analysis, Labeling 세분화를 통해 필요 소리 데이터 종류 파악 및 추가 수집
 - 모델과 관련하여 성능이 더 밀접한 건 Parameter의 수
- 향후 발전 가능성 및 보완 사항
 - 실험 내 성능과 데모의 성능의 불일치
 - Feature를 가져오는 단계에서 일치를 시켜주는 프로세스에 대한 연구 필요
 - 실제 적용을 위한 추가적인 데이터 수집



Our Team



신현경 투빅스 9기
덕성여자대학교 정보통계학과



신훈철 투빅스 10기
홍익대학교 산업공학과



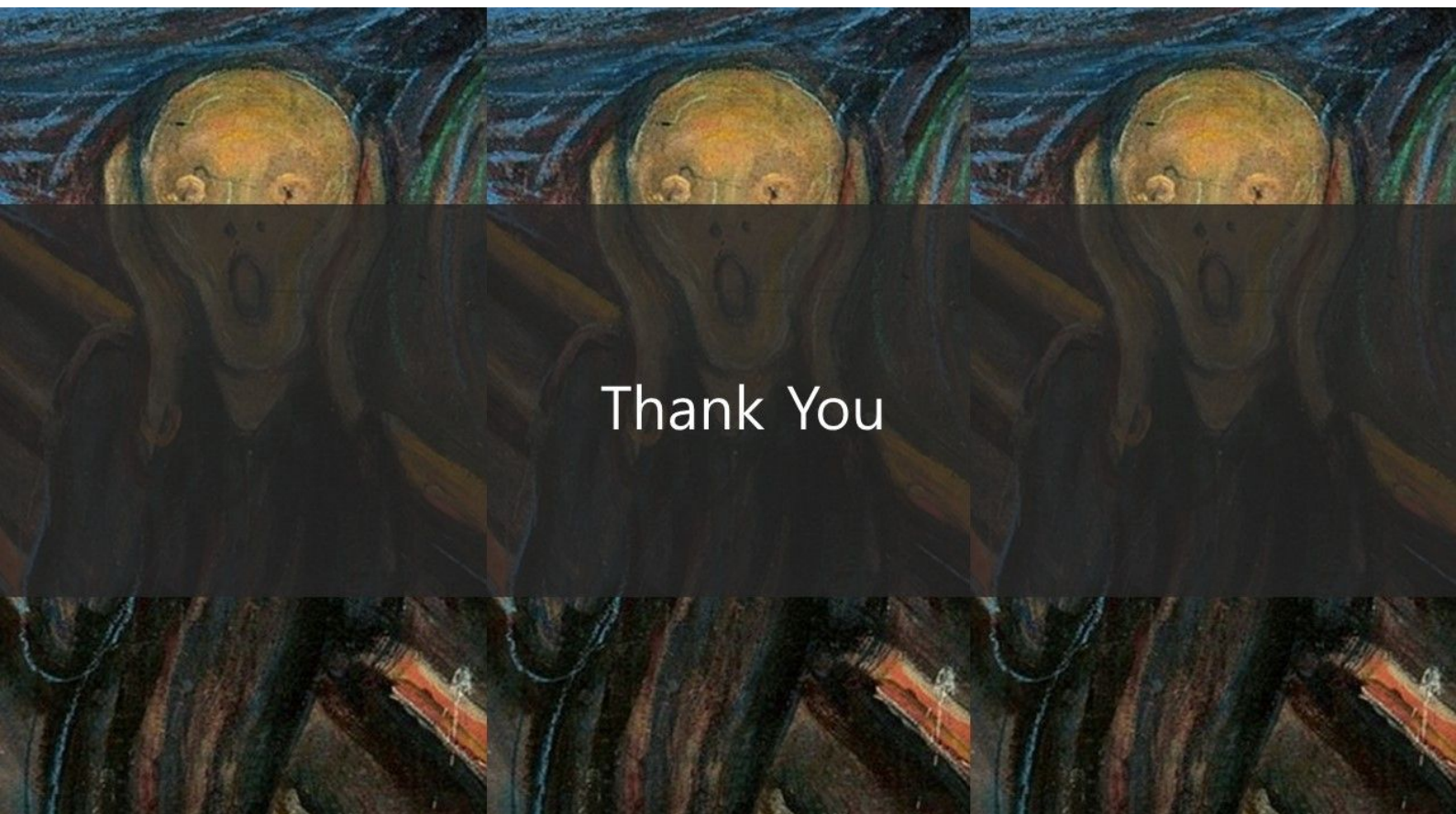
유기운 투빅스 11기
서울대학교 식품생명공학과



임채빈 투빅스 11기
가천대학교 응용통계학과



최세영 투빅스 10기
덕성여자대학교 컴퓨터공학과



Thank You