

비정형 빅데이터 분석의 응용과 실습

Week-04. Audio Data

서중원 2020.09.26

텍스트는 String
이미지는 Matrix
비디오는 연속된 이미지
오디오는...?

소리의 구성성분

먼저 자연계에서의 소리란?

- 소리의 3요소
- 주파수 : 음의 높낮이, Hz
 - 진동수가 높은 음을 고음
 - 진동수가 낮은 음을 저음
 - 돌고래: 초음파
 - 호랑이: 저주파
- 진폭 : 음의 세기, dB
- 파형 : 음색

| 물리량 (physical quantity) | 심리량(subjective quantity) | | |
|----------------------------|--------------------------|-------------------|----------------|
| | 소리 크기 (loudness) | 소리 높낮이 (pitch) | 음색 (timbre) |
| 음압(pressure) | *** | * | * |
| 주파수(frequency) | ** | *** | ** |
| 스펙트럼(spectrum) | * | * | *** |
| 포락선(envelope) | * | * | ** |
| 지속시간(duration) | * | * | * |

(상관정도: *)

소리의 3요소



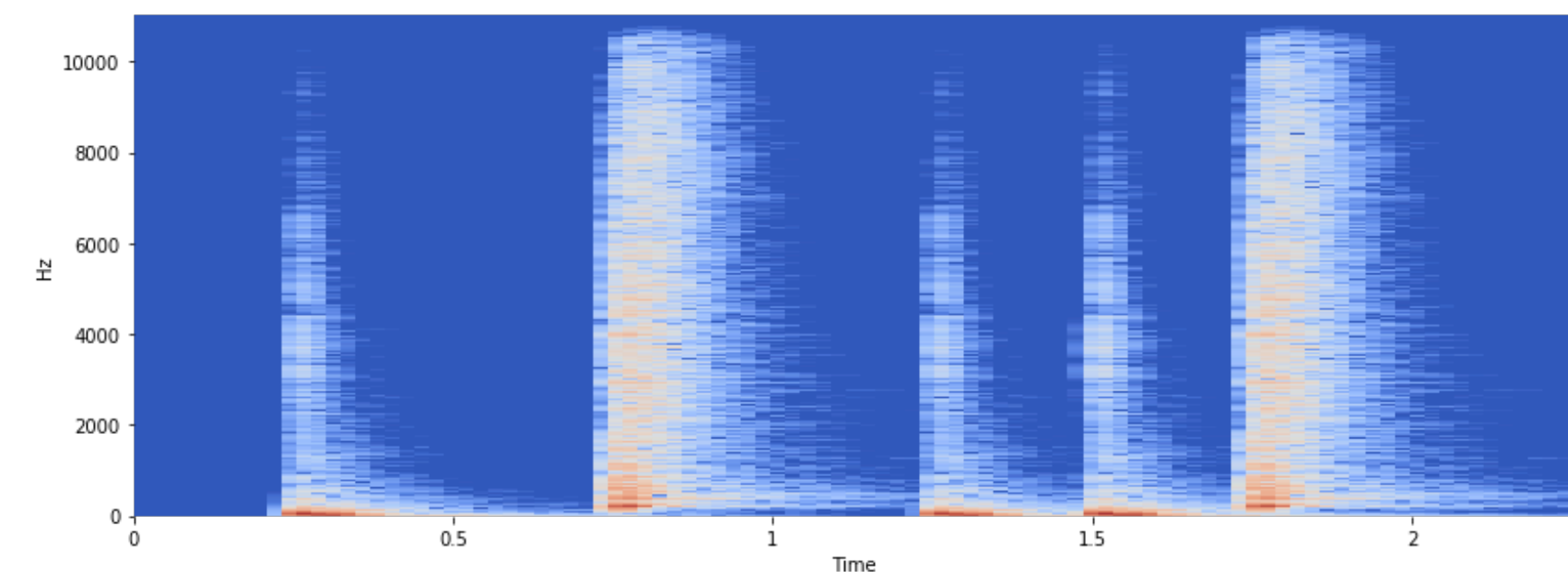
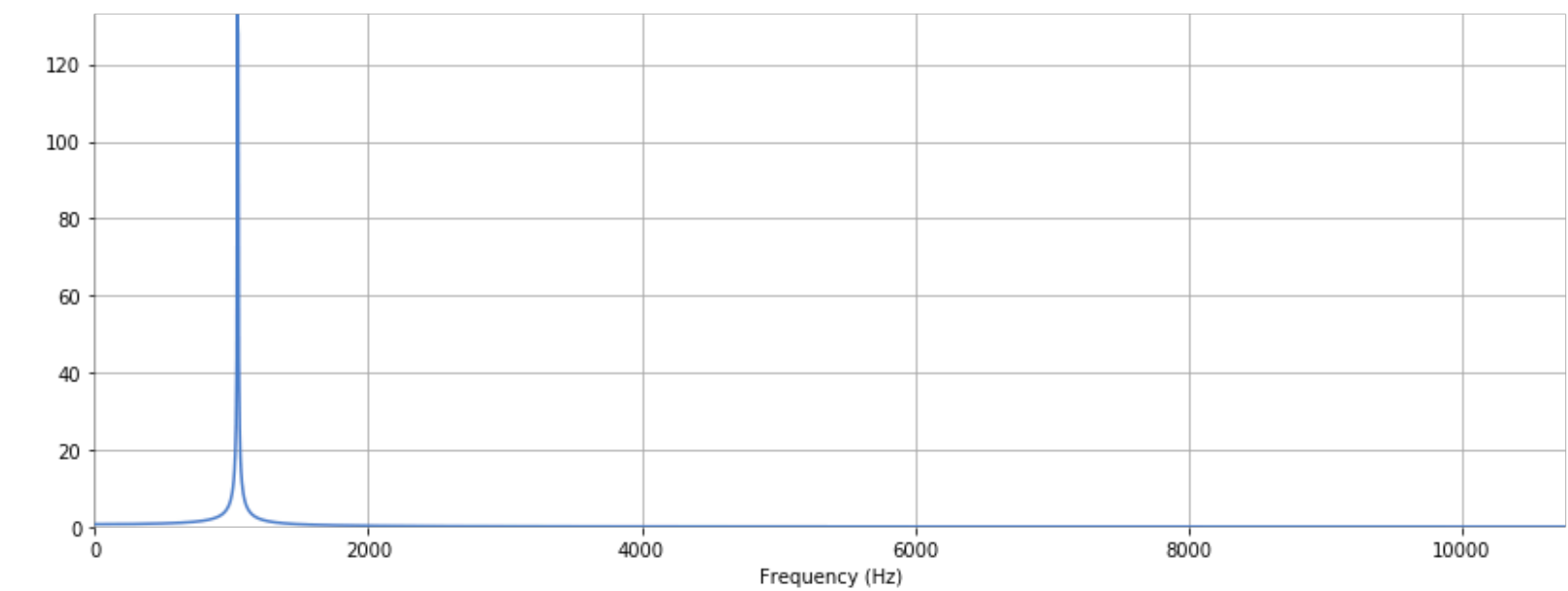
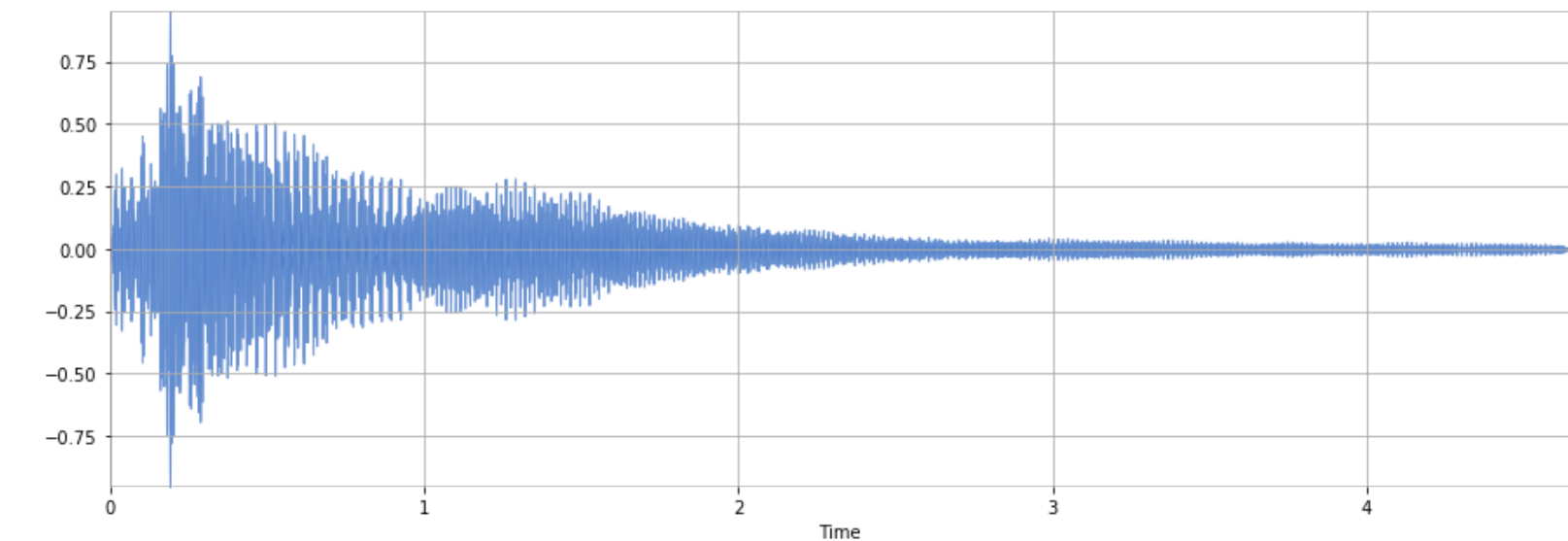
* KISTI의 과학향기, 호랑이 울음 소리에 기가 죽는 이유는? <한겨레> 2006-03-10 17:35

** http://www.ktword.co.kr/abbr_view.php?nav=&m_temp1=5066&id=1005

오디오의 표현방법

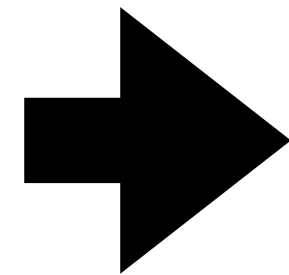
오디오 데이터를 컴퓨터가 이해할 수 있도록 하려면?

- 방법 1
 - 시간-세기 그래프
- 방법 2
 - 주파수 그래프
- 방법 3
 - 주파수의 강도와-시간 그래프

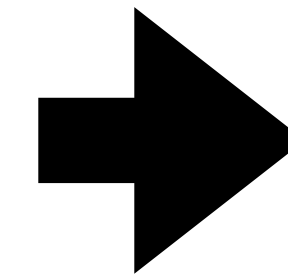


오디오 데이터

오디오를 벡터로 벡터를 오디오로



[12,3004, 300, 0,0,300 ...]



우리가 관심가져야 할 영역


오디오 포맷의 종류

자주 사용하는 포맷

- **mp3**
 - MPEG Layer 3 오디오 파일. 오늘날 가장 널리 쓰이는 오디오 파일 포맷
- **webm**
 - HTML5 비디오 용으로 제작 된 로열티없는 형식
- **wav**
 - 윈도 PC에서 주로 쓰이는 표준 오디오 파일 컨테이너
 - 보통 비압축 방식의 CD급 품질 오디오 파일을 저장하기 위해 사용
 - 비압축 방식이므로 필연적으로 파일의 크기가 클 수밖에 없다

그렇다면 음성인식은 어떤 원리로..?


텐서플로우 음성인식 챌린지!

 Featured Prediction Competition

TensorFlow Speech Recognition Challenge

Can you build an algorithm that understands simple speech commands?

\$25,000
Prize Money

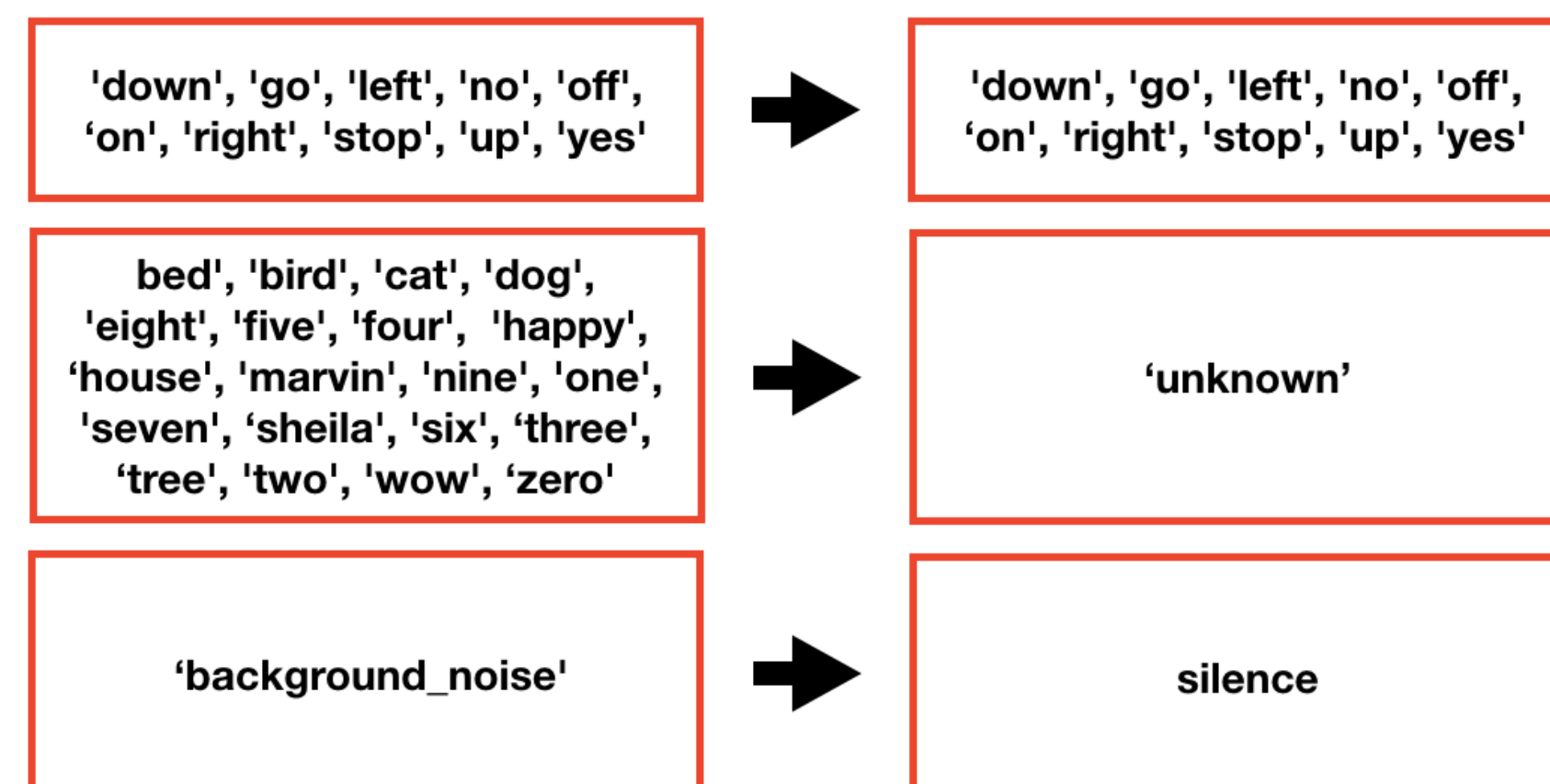
 Google Brain · 1,315 teams · a year ago

<https://www.kaggle.com/c/tensorflow-speech-recognition-challenge>

챌린지 소개

문제와 데이터

- Task 타입
 - 지도학습-분류문제
- 데이터셋
 - 오디오 파일: WAV format
 - 훈련데이터
 - 라벨링 된 데이터: 64,727개
 - 라벨링 안된 데이터: 158,538
 - Label의 갯수 : 31개
 - 맞춰야 하는 클래스 :12개



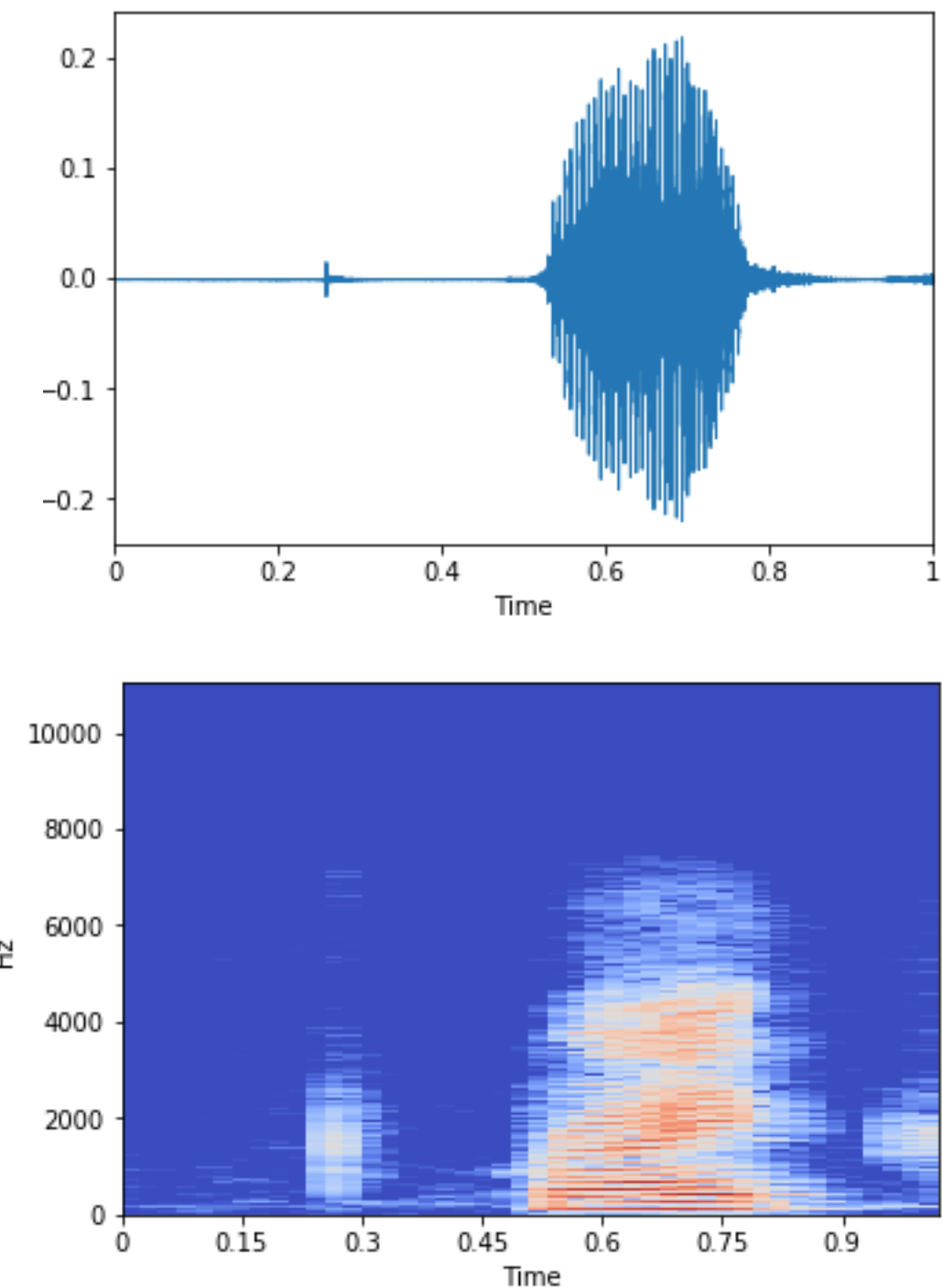
| | | | | | |
|-------------------------|--------------|--------------|-------------|---------------|--------------|
| Right | Eight | Cat | Tree | Bed | Happy |
| 2367 | 2352 | 1733 | 1733 | 1713 | 1742 |
| No | Wow | Nine | Left | Stop | Three |
| 2375 | 1745 | 2364 | 2353 | 2380 | 2356 |
| Bird | Zero | Seven | Up | Marvin | Two |
| 1731 | 2376 | 2377 | 2375 | 1746 | 2373 |
| Six | Yes | On | Five | Off | Four |
| 2369 | 2377 | 2367 | 2357 | 2357 | 2372 |
| Dog | One | Down | Go | Sheila | House |
| 1746 | 2370 | 2359 | 2372 | 1734 | 1750 |
| Background_noise | | | | | |
| 6 | | | | | |

Table 1: The number of training audio files per class

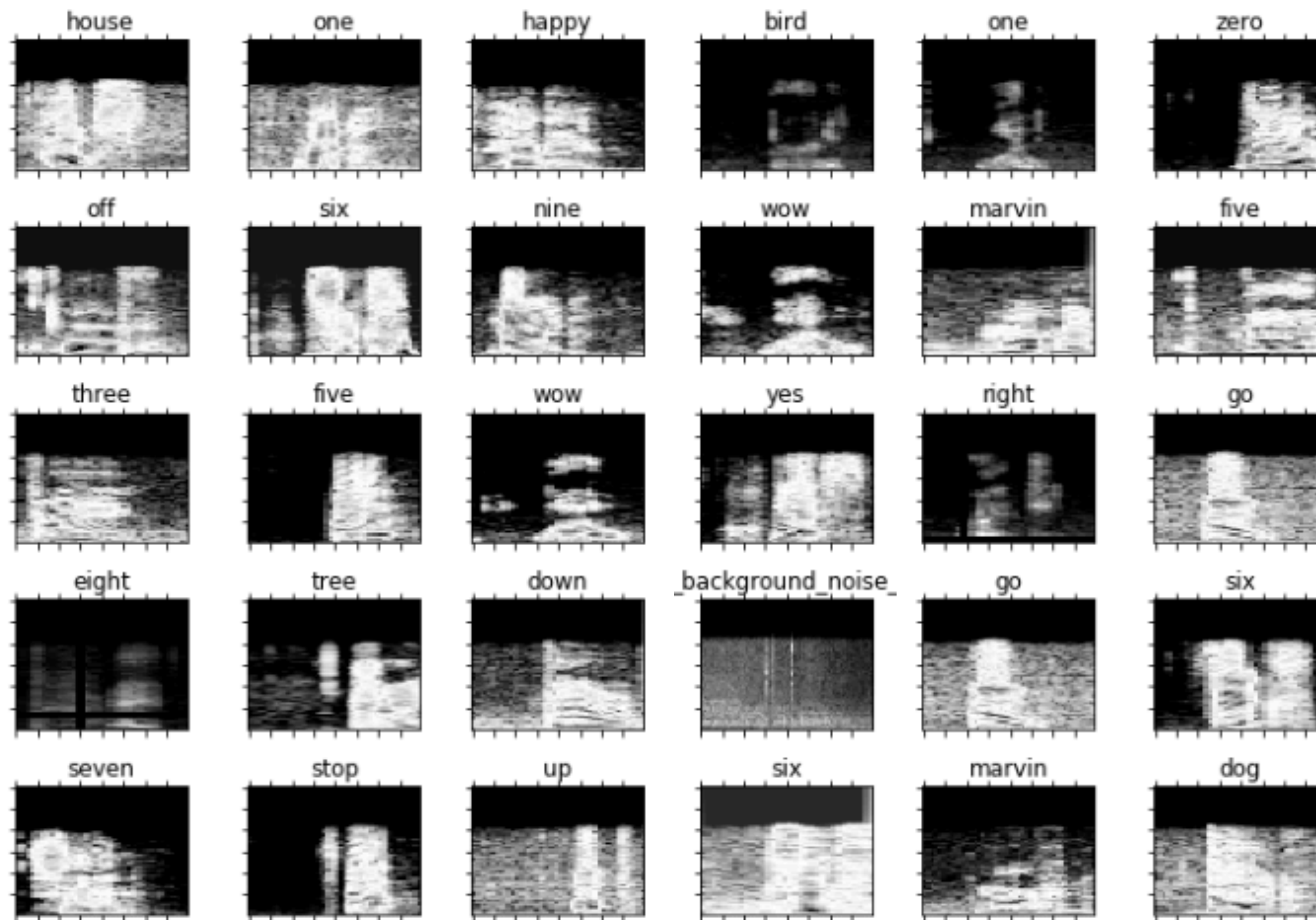
전처리 과정

오디오를 Matrix로

- 아직까지도 오디오 표현(Representation)에 대한 정답은 없는 상태
- Waveplot
 - 진폭(Amplitude)를 시간순으로 나열
- STFT(Short-time Fourier-Transform)
 - Amplitude로만 표현을 하면 Speech를 표현하지 못한다는 관점에서 많이 사용
 - 같은 진폭의 발화는 다 같은 소리인가?
 - 푸리에 변환을 작은 시간 단위로 쪼개서 진행
 - 주파수와 시간을 동시에 표현할 수 있다는 점이 장점



어디서 많이 본것 같지않나요?

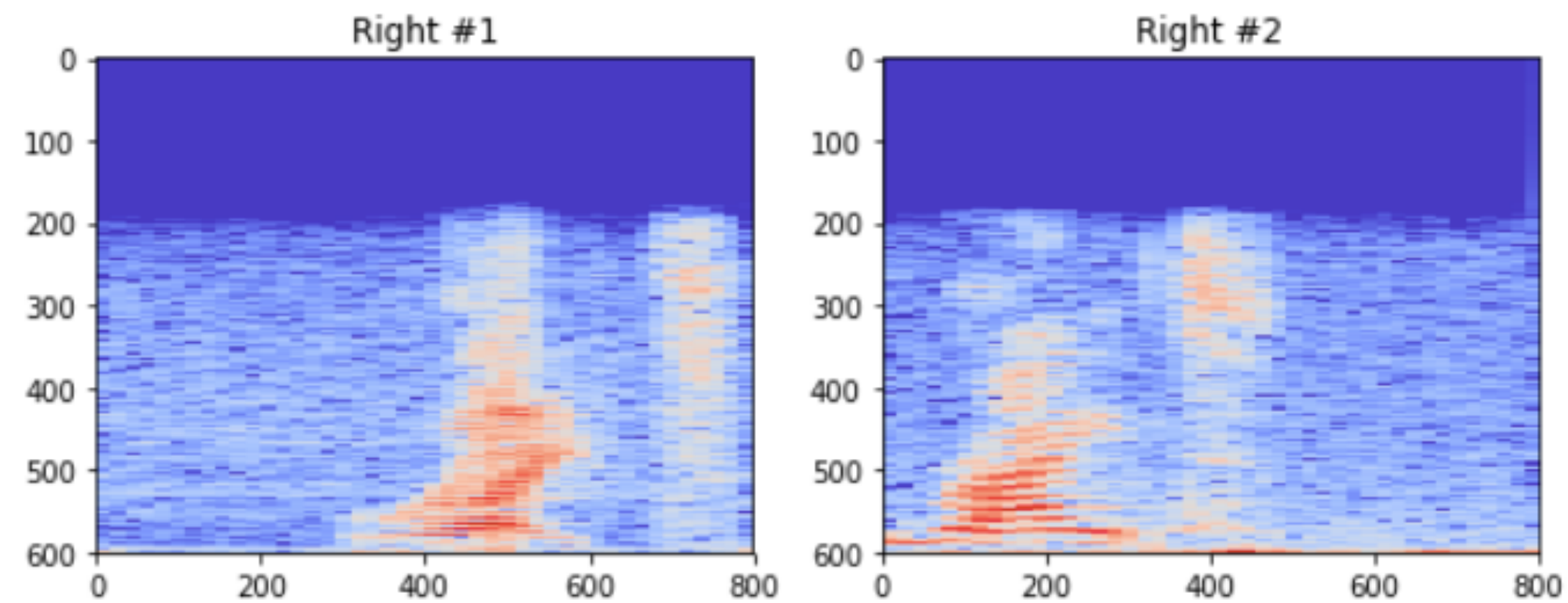




모델 선정

후보 1. CNN

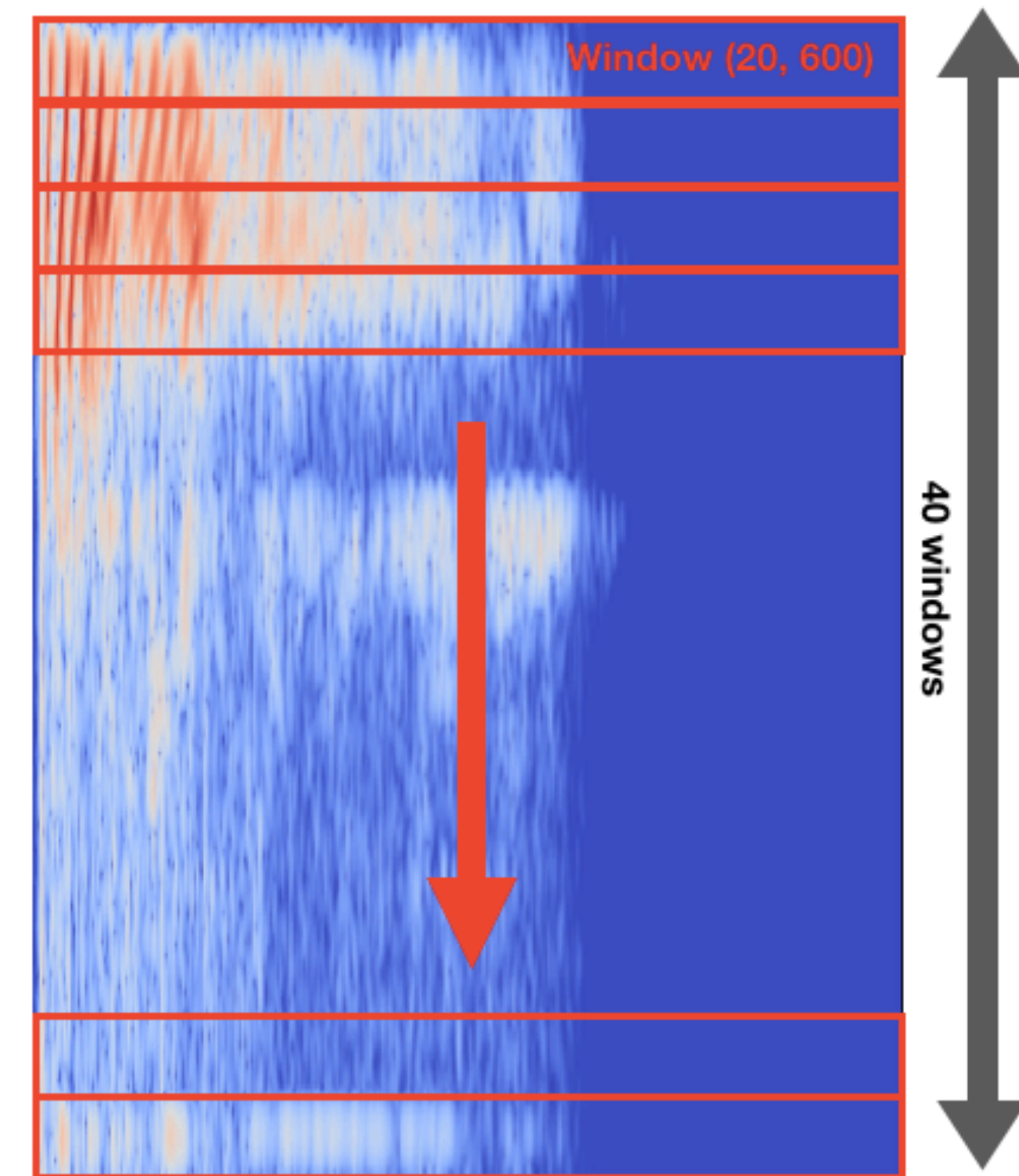
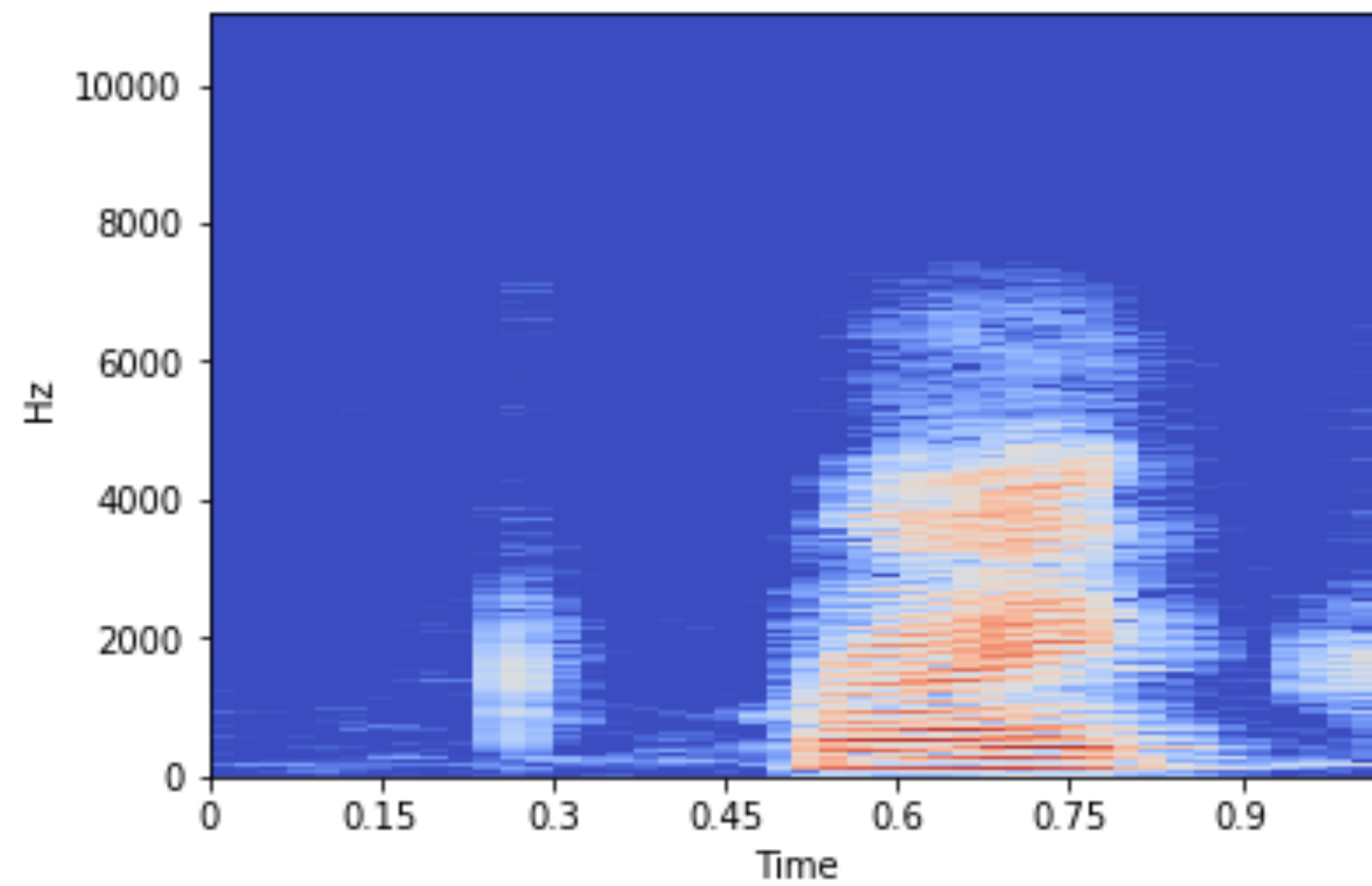
- 같은 단어를 말해도 시작점이 다르다
- 같은 단어도 화자의 속도가 다르다



모델 선정

후보 2. RNN

- 결국에 음성도 시계열 데이터기 때문에, RNN을 쓰는게 가장 적합하다



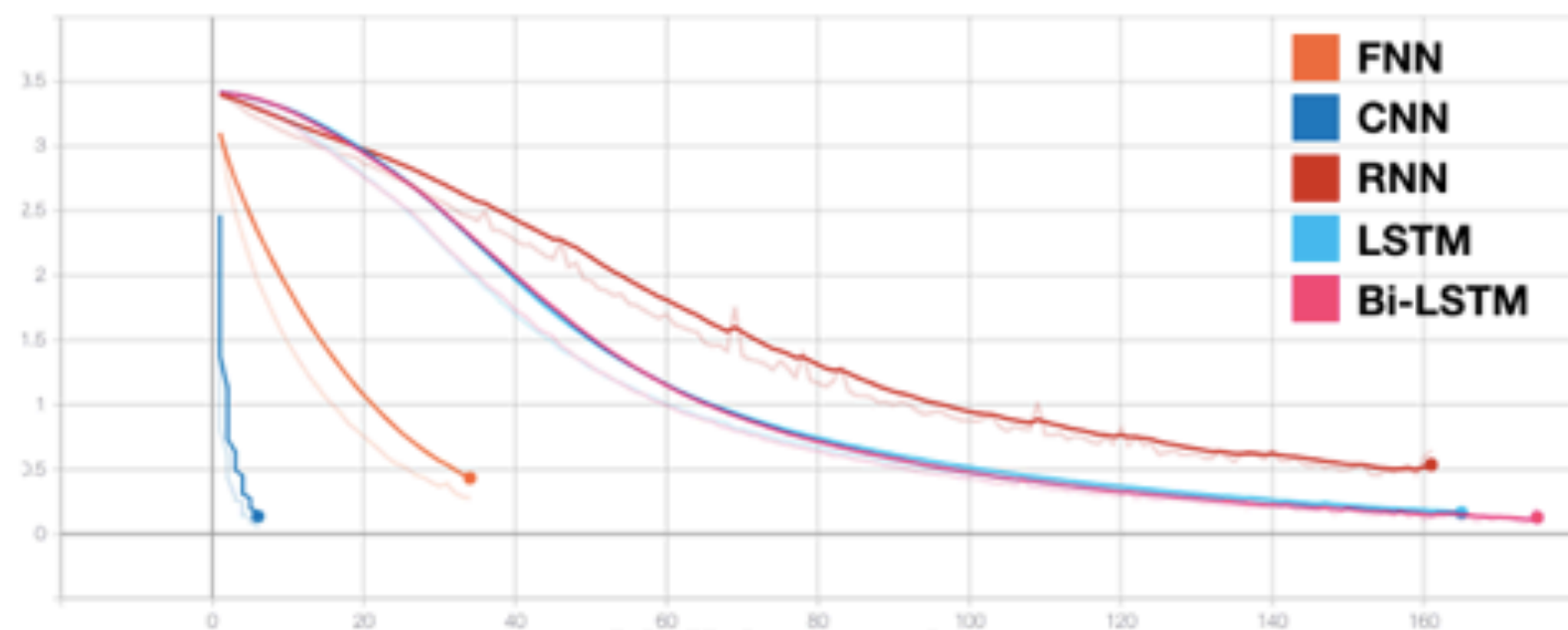
성능 평가

누가 가장 잘 했을까?

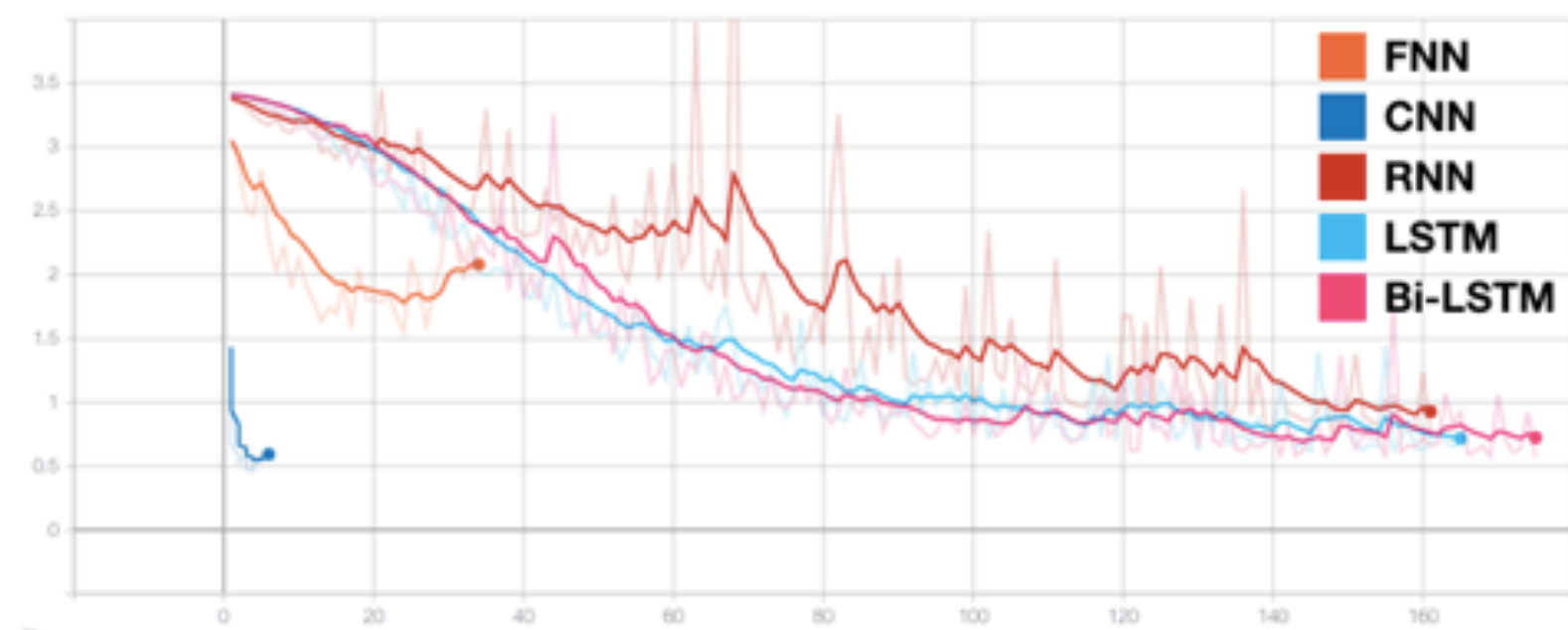
Accuracy



Train Loss



Validation Loss



성능평가

속도와 정확도의 Trade-Off

- BI-LSTM이 가장 좋은 성능을 보이지만, 훈련시간이 CNN에 비해 10배가 넘게 걸리기 때문에 CNN으로 진행
 - 그 중에서 잘 알려진, Resnet으로 진행

| | FNN | CNN | RNN | LSTM | BILSTM |
|-----------------------|------------|-----------|------|------|-------------|
| Training Acc. (%) | 58 | 87 | 80 | 82 | 85 |
| Kaggle Acc. (%) | 48.7 | 65.2 | 58.5 | 63.7 | 65.7 |
| # of Epochs | 34 | 6 | 161 | 165 | 175 |
| Training Time (hours) | 1.5 | 1.6 | 14 | 15 | 17.7 |

Table 3: Result of candidate models

모델 최적화만으로 진행 전처리는 나중에 다시!

- 78%의 정확도!

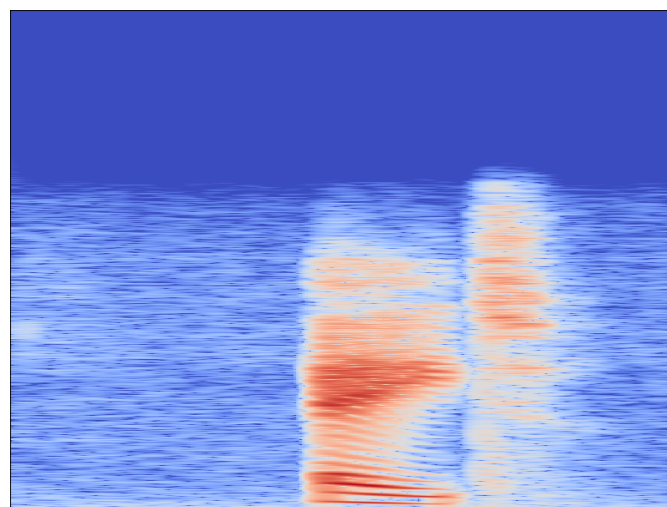
| | | | |
|--|---------|---------|--------------------------|
| submission_resnet.csv 3 days ago by Jungwon Seo RESNET no background | 0.78832 | 0.78042 | <input type="checkbox"/> |
|--|---------|---------|--------------------------|

처음 부터 다시 점검

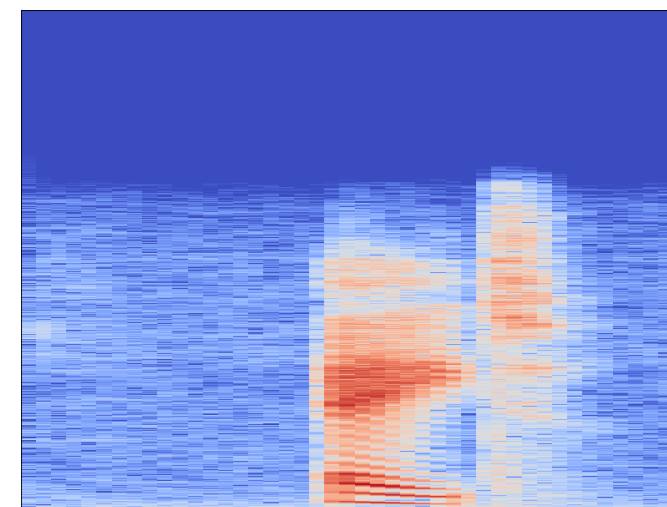
전처리는 잘 되었나?

- Silence에 해당하는 데이터셋이 뽑히지 않았다.
- 데이터셋의 균형을 맞추자
 - 데이터 Augmentation을 통해서

| | Imbalanced Dataset | Balanced Dataset |
|---------|--------------------|------------------|
| Down | 4576 | 5731 |
| Go | 7093 | 6425 |
| Left | 5921 | 7752 |
| No | 9015 | 6805 |
| Off | 5907 | 6541 |
| On | 10443 | 7125 |
| Right | 6136 | 5551 |
| Stop | 6036 | 6314 |
| Up | 15450 | 8999 |
| Yes | 5562 | 5071 |
| Unknown | 82399 | 84645 |
| Silence | 0 | 7579 |



Original



Different hop/window

성능평가

다시 확인!!

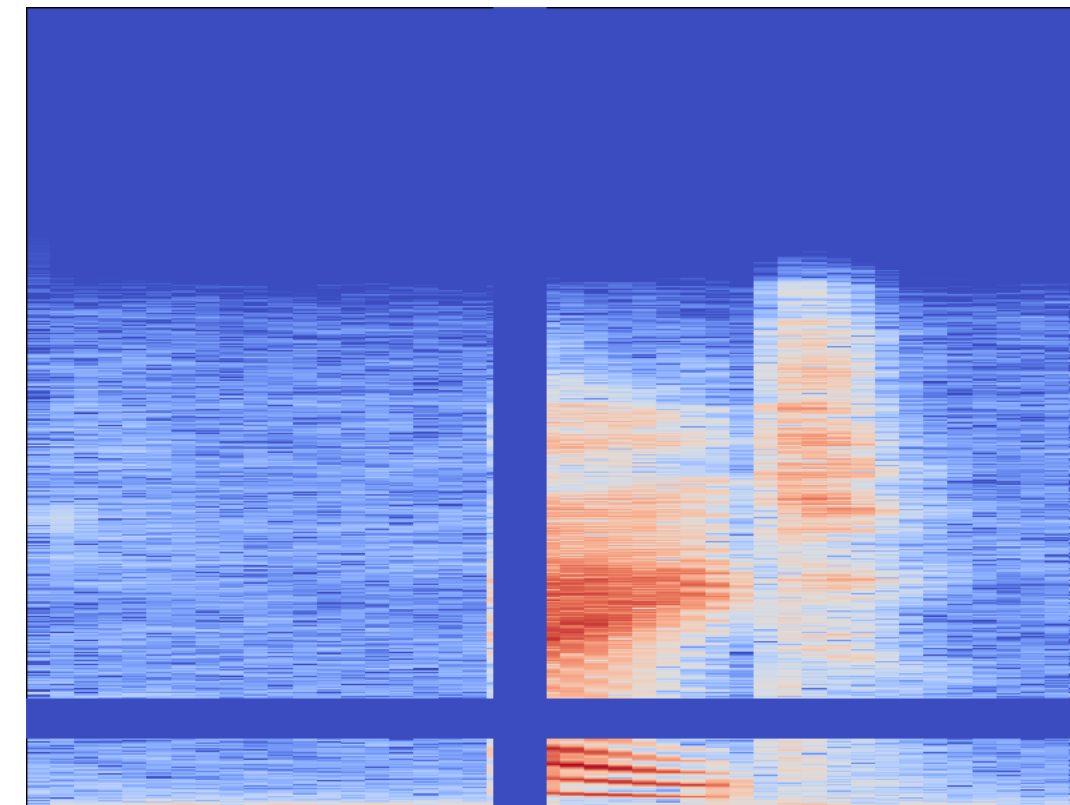
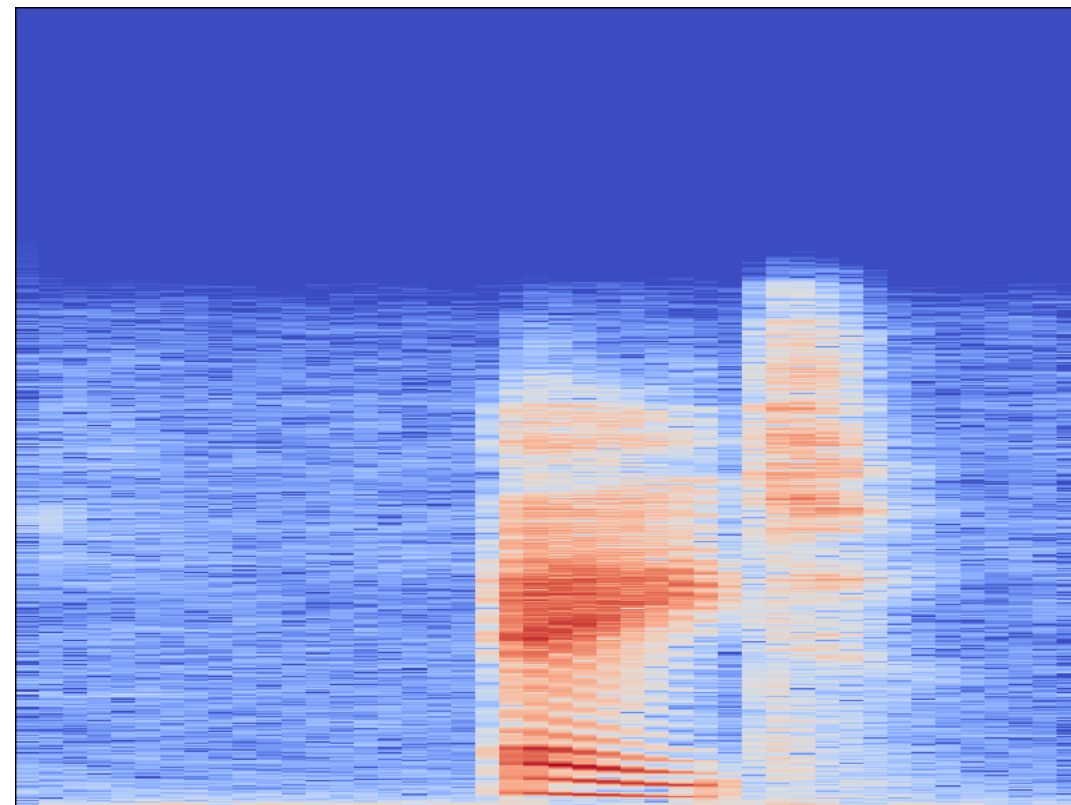
- 0.001~0.002의 실망스러운 증가

| | | | |
|--|---------|---------|--------------------------|
| submission_resnet.csv 2 days ago by Jungwon Seo RESNET over training | 0.79278 | 0.78125 | <input type="checkbox"/> |
| submission_resnet.csv 2 days ago by Jungwon Seo resnet 18 with balanced data | 0.78855 | 0.78234 | <input type="checkbox"/> |

더 많은 데이터

데이터를 더 확보하자!

- 새로운 데이터 강화 방법이 나왔다던데..?



SpecAugment: A New Data Augmentation Method for
Automatic Speech Recognition

Monday, April 22, 2019

Posted by Daniel S. Park, AI Resident and William Chan, Research Scientist

최종결과 과연!

- 뭔가 다른 문제가 있던 것 같다

| | | | |
|---|---------|---------|--------------------------|
| submission_resnet_new5.csv a few seconds ago by Jungwon Seo PLEASE..... | 0.78714 | 0.77549 | <input type="checkbox"/> |
|---|---------|---------|--------------------------|

과제 리뷰

이 과제를 하면서의 실수

- 딥러닝에 너무 신이난 나머지, 훈련만 열심히 시켰다.
- 전처리에 더 집중을 했어야 했다.
- 무작위 식으로 계속 하이퍼 파라미터만 바꿔가며 테스트 하지 말았어야 했다.
- 기본적인 EDA 부터 제대로 다시 했어야 했다.

E.O.D