Project 1 Report

2021327348 이철환

1. HMM train and validation

HMM을 활용한 음성 인식 모델을 학습하고 여러 validation 데이터를 활용하여 성능을 테스트한다.

HMM training

```
labels_train = {
    '11jeonghy',
    'Dandyst',
    'InkooJeon',
    'YouYeNa',
    # 'deokkyukwon',
    # 'ohjihyeon',
    'son',
}
```

왼쪽 레이블에 해당하는 데이터를 학습 데이터로 활용하였다. 학습 데이터는 filtering 및 노이즈 추가 없이 그대로 활용하여 모델을 생성한다.

모델 입력은 6개의 MFCC를 활용한다. 20ms 크기(160 samples) window, 10ms 크기 hop으로 windowing하여 (n, 6) 크기의 MFCC를 생성하여 모델에 입력한다.

학습된 모델은 각 레이블과의 유사도를 출력하는 10개의 모델로 이루어져 있다. 각 모델의 추론 결과를 토대로 가장 유사도가 높은 레이블을 최종 추론 결과로 선택한다.

Validation

Validation 데이터는 다음 2명의 데이터를 활용하였다. 각 데이터에 대한 validation 결과는 다음과 같다.

```
labels_val = {
    'chlee',
    'do',
    # 'kyeong',
}
```

```
train : accuracy = 47.0
wbnSNR-10 :accuracy = 9.5
org :accuracy = 44.0
wbnSNR10 :accuracy = 29.0
nbnSNR0 :accuracy = 24.0
wbnSNR0 :accuracy = 16.5
nbnSNR-10 :accuracy = 10.0
nbnSNR10 :accuracy = 35.5
```

동일한 SNR에서는 nbn이 wbn에 비해 높은 정확도를 보였고, SNR이 작아질수록 정확도가 크게 낮아짐을 확인할 수 있었다.

2. HMM test

EPD를 활용한 테스트 데이터 분할

Waveform 데이터의 에너지를 활용하여 음성이 나타나는 구간을 탐색하고, 이를 토대로 분할되지 않은 테스트 데이터를 분할하여 모델에 입력하여 정확도를 측정한다.

$\sqrt{\frac{1}{N_S}\sum x^2(t)}$ 로 프레임별 총 에너지의 평균을 구하고, 에너지의 크기가 threshold를 넘으면 음성 구간으로

판단한다. 프레임간 간격이 좁으면 하나의 음성 구간으로 판정한다.

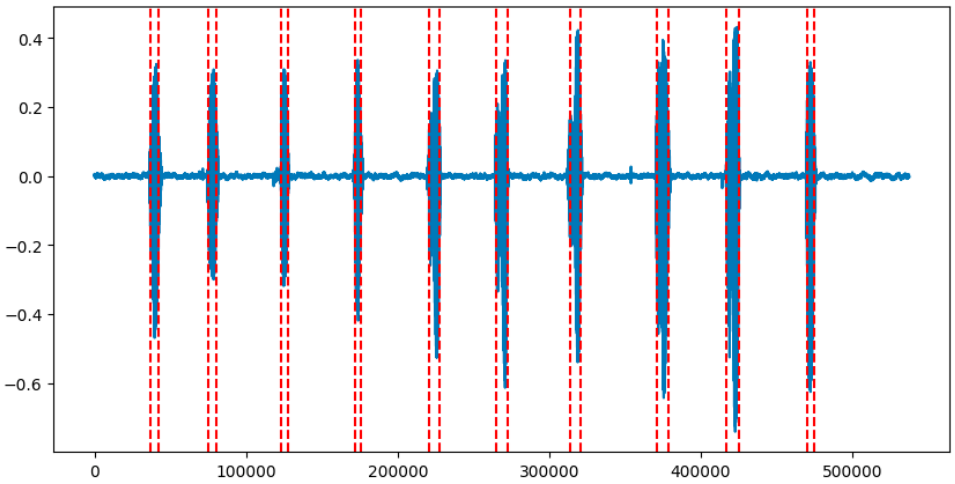
EPD 함수의 인자는 다음과 같다.

```
find_endpoints(x, frame_t=0.002, energy_thd=0.1, t_gap_thd=100):
```

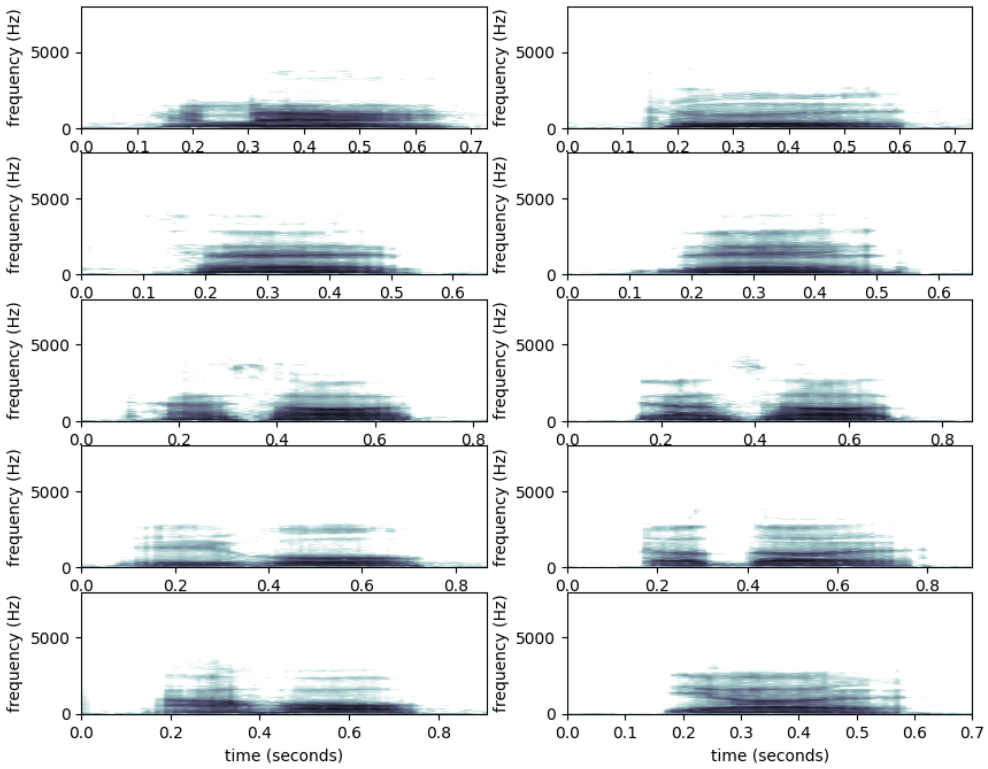
- frame_t: 에너지를 계산할 프레임 크기
- energy_thd: 각 프레임을 음성 구간으로 판정할 최소 평균 에너지
- t_gap_thd: 음성 구간 사이 에너지가 낮은 구간 발생 시 하나의 음성 구간으로 판별할 최대 프레임 간격

EPD 수행 결과

Clean data: energy_thd=0.1, t_gap_thd=100



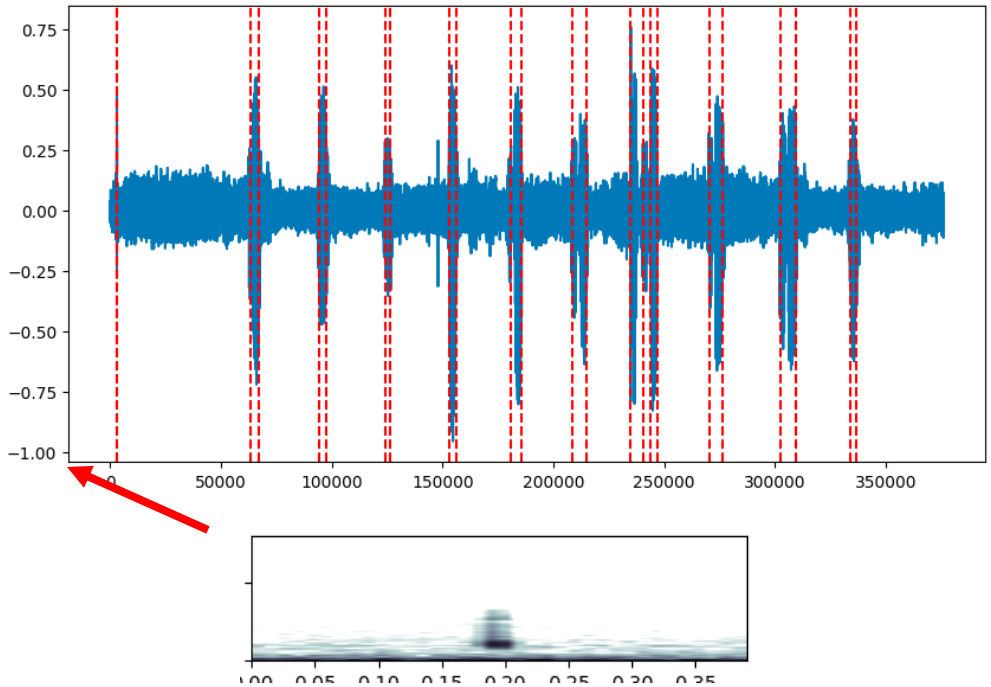
EPD 구간



분할한 데이터의 spectrogram (차례대로 1~10까지 음성 구간)

Clean 데이터를 대상으로는 10개의 파일 모두 위와 같이 10개의 음성 구간을 거의 정확히 구분하였다.

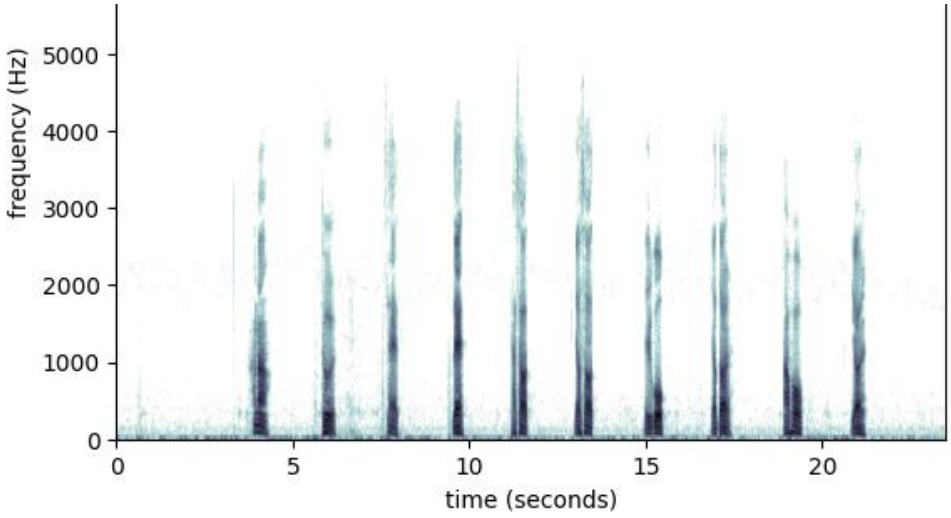
Noisy data: SNR0, energy_thd=0.2, t_gap_thd=100



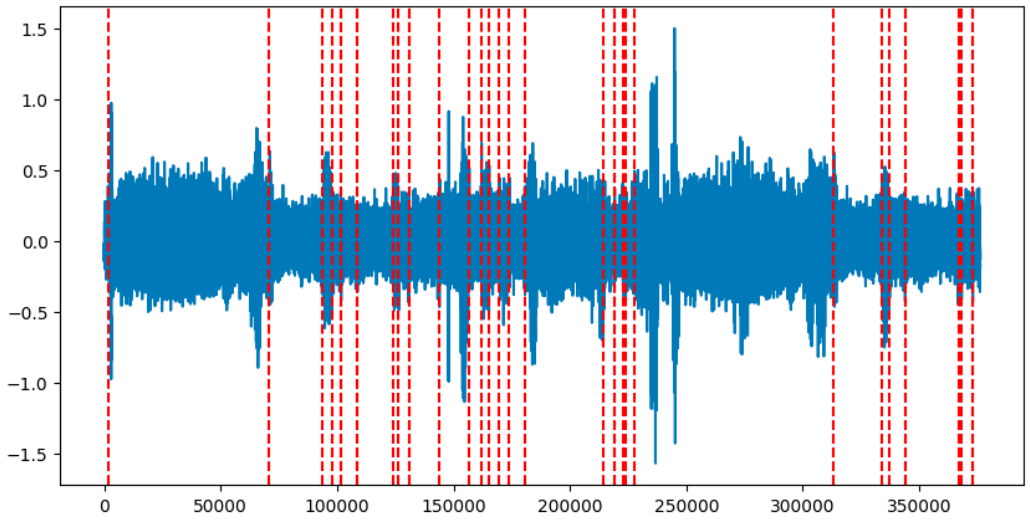
Noisy 데이터에서 EPD

노이즈가 발생한 데이터의 경우 에너지가 높은 노이즈를 음성으로 판별하는 경우가 발생하였다. 데이터에

따라서는 threshold 값 조정을 통해 구분할 수 있는 경우도 있었으나, 일반화하기는 어려웠다. Threshold 를 더 높게 설정할 경우 두 음절로 이루어진 발음의 절반이 소실되는 경우가 발생하였다.



다만 전체 데이터를 봤을 때 잡음이 발생하는 스펙트럼 구간과 음성이 갖는 스펙트럼 구간이 차이가 많이 나므로, 음역대별 에너지를 토대로 EPD를 수행한다면 구분 가능할 것으로 생각된다.



EPD SNR-10

SNR-10 데이터에서는 EPD가 거의 동작하지 않았다. 음성 구간을 합치기 위해 설정한 threshold가 잡음으로 채워져 구간이 거의 구분되지 않았고, 육안으로도 음성과 노이즈를 구분하기 힘들었다. 스펙트럼에서도 음성 구간을 가리는 노이즈가 많이 발견되어 EPD를 통한 구분은 힘들 것으로 보인다.

HMM test

노이즈가 포함된 데이터에서는 EPD를 통한 음성 구간 분할 및 레이블링이 어려워 Clean 데이터에서 구한 EPD를 토대로 전체 데이터를 분할하여 테스트 데이터셋을 생성하였다. 추론 결과는 다음과 같다.

```
org : accuracy = 26.0
nbnSNR-10 : accuracy = 20.0
nbnSNR0 : accuracy = 20.0
nbnSNR10 : accuracy = 28.0
wbnSNR-10 : accuracy = 10.0
wbnSNR0 : accuracy = 10.0
wbnSNR10 : accuracy = 17.0
```

육안으로는 음성 구간이 잘 구분된 것으로 보였으나, 정확도는 validation 데이터에 비해 비교적 낮게 나타났다. Validation과 동일하게 SNR이 낮고 wbn 잡음이 있을 때 정확도가 낮게 나타났다.

3. Winer filtering

Filtering 부분은 아직 이론 및 코드 이해가 부족하여 적용하지 못했습니다. 다음과 같이 구현 계획을 잡았습니다.

- Validation data: EPD를 적용해 음성 구간을 구하고, 이외 구간을 노이즈로 보고 필터를 생성해 잡음 제거
- Test data: EPD를 활용해 구한 잡음 구간으로 필터 생성 -> 전체 구간에 대해 필터링 -> 필터링된 데이터에 Threshold를 변경하여 EPD를 수행하여 음성 구간 재분할

SNR0 까지는 잡음 제거를 통한 분류 성능 향상을 기대할 수 있을 것으로 예상합니다.