영상 테스트

n_fft와 hop_length 파라미터 조정이 큰 영향이 없음을 알고, 학습데이터는 다양한 파라미터로 늘려서 만들어 학습시키고, predict용 데이터는 파라미터별로 테스트해보고 제일 결과가 좋은 파라미터로 기록했습니다.

음성 종류 (목소리, 노래, 생활 잡음)와 화자 성별(여성, 남성) 구분하는 모델 2가지의 테스트 결과까지 보여드립니다.

dataset 준비

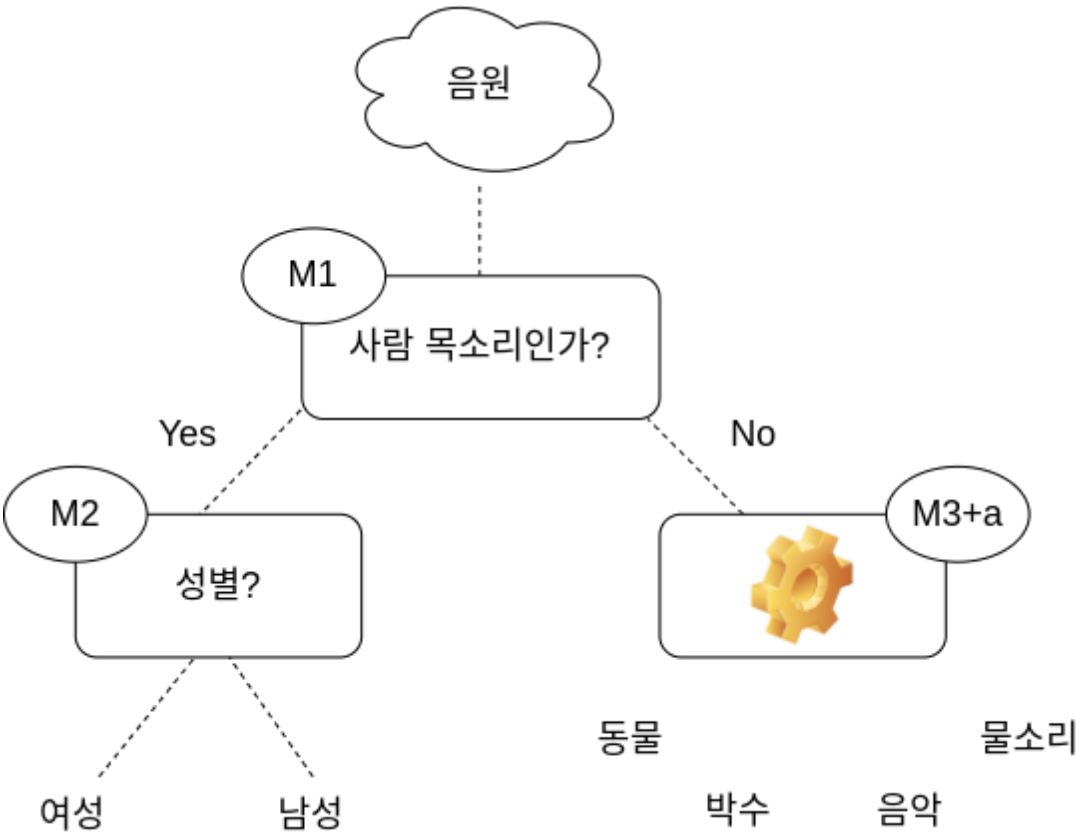
- n_fft 512, 1024, 2048, 4096 - hop_length는 n_fft*0.25로 총 4가지 파라미터 조합으로 생성
- 이전에 파라미터 테스트하면서 만들어둔 일부에 이어 가지고 있는 전체 데이터 생성을 위해 풀가동!
- 병렬로 4개 동시에 돌려서 하루 정도 소요

	path	n_fft	hop_length	duration	chroma_stft_mean	chroma_stft_var	stft1_mean	stft2_mean	stft3_mean
0	../../sounds/human_voice/read_men/nVToV-zn_C4....	512	64.0	2.730023	0.736829	0.011109	0.631880	0.710417	0.801867
1	../../sounds/human_voice/read_men/nVToV-zn_C4....	512	128.0	2.730023	0.736613	0.011078	0.634239	0.713801	0.804068
2	../../sounds/human_voice/read_men/nVToV-zn_C4....	512	256.0	2.730023	0.743991	0.010503	0.642011	0.718064	0.807828
3	../../sounds/human_voice/read_men/nVToV-zn_C4....	1024	128.0	2.730023	0.612632	0.000370	0.633576	0.634752	0.627650
4	../../sounds/human_voice/read_men/nVToV-zn_C4....	1024	256.0	2.730023	0.613040	0.000373	0.634273	0.635422	0.628011
...	...	...	...	...	...	...	...	...	...
98627	../../sounds/human_voice/read_women/4_0220.wav	4096	1024.0	1.904036	0.451054	0.001506	0.456886	0.428688	0.510192
98628	../../sounds/song/Soul_Pop_Moody_please_95BGM/...	512	128.0	3.000000	0.690494	0.007287	0.562173	0.634749	0.724480
98629	../../sounds/song/Soul_Pop_Moody_please_95BGM/...	1024	256.0	3.000000	0.559983	0.011119	0.441183	0.503367	0.753819
98630	../../sounds/song/Soul_Pop_Moody_please_95BGM/...	2048	512.0	3.000000	0.439074	0.017911	0.319693	0.345197	0.698272
98631	../../sounds/song/Soul_Pop_Moody_please_95BGM/...	4096	1024.0	3.000000	0.372416	0.018779	0.270796	0.216808	0.553799

132873 rows × 87 columns

Pic-ensemble 모델 계획

- 드라마픽에 들어가는 앙상블 모델이라는 뜻에서 이름을 붙여주었습니다
- 아직은 간단하지만..라벨 추가에 따라 무럭무럭 성장할 모델을 기원하면서..



모델 학습

- train-test셋 사전 분리 세팅 (파라미터값이 다른 두 음원이 train-test에 따로 들어가지 않도록 방지)
- n_fft별로 train-test셋을 분리해서도 학습해보고, 전체 데이터셋으로 학습해보았는데 성능이 차이가 없었습니다.
- 4가지 파라미터의 데이터셋을 모두 포함해서 학습시켰습니다.
- 전체 데이터로 학습시킨 뒤, test셋은 특정 n_fft별로 테스트진행
    - 추후 실제 predict 진행시 frequency데이터는 정해진 파라미터로 일괄 추출하기 때문에 predict 용 파라미터값을 default값으로 진행해도 될지 최종으로 정하기 위해서

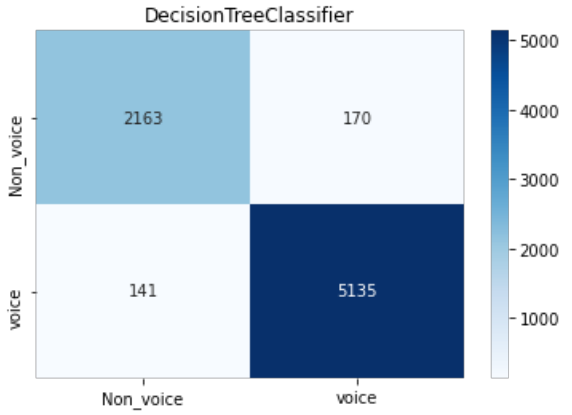
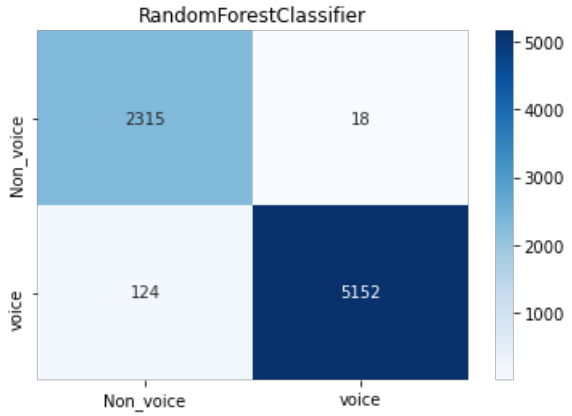
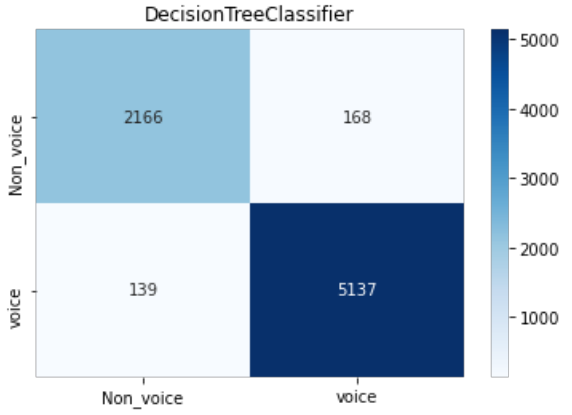
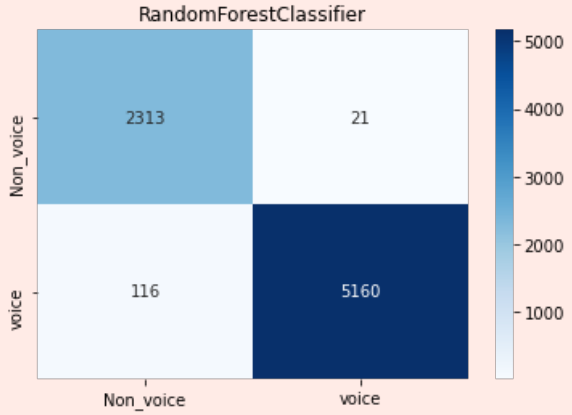
M1 사람 목소리 인가?

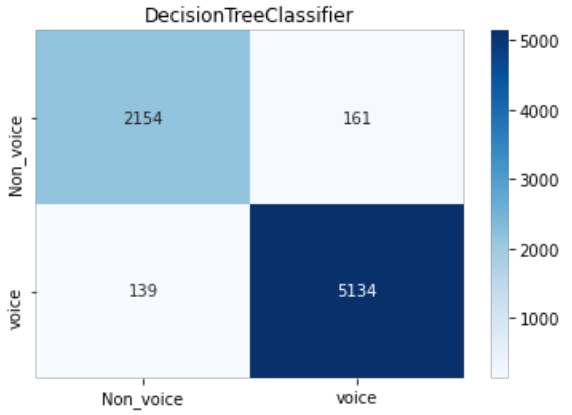
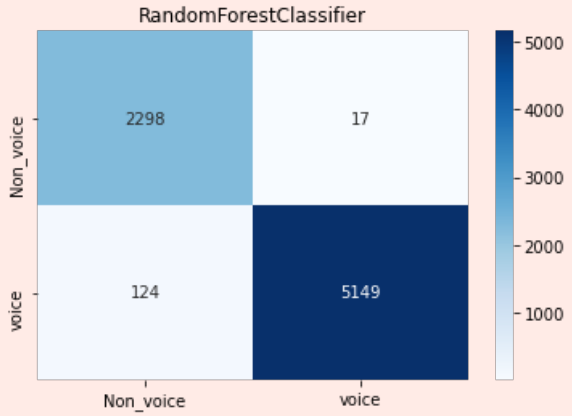
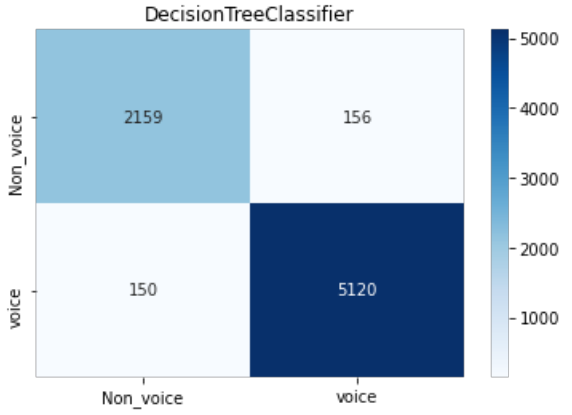
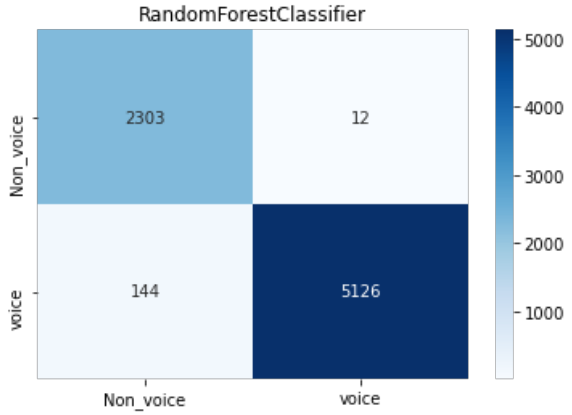
Train - test

- Train셋
    - 총 7만개
    - human_voice: 사람 목소리(여자+남자) - 4.7만
    - Non_human_voice: 기타 생활잡음+자연소리+노래소리 (동물은 우선 배제했습니다) - 2.3만
- test셋 - 파라미터별 결과 확인 → Default값 2048로 무방할 것으로 보이며, **1024로 진행하기로 결정**!

nfft	acc	recall	f1
1024	0.981997	0.984508	0.979068
2048	0.981418	0.984570	0.978364
512	0.981338	0.984391	0.978325
4096	0.979433	0.983746	0.976126

n_fft - hop_length	DT	RF

512 - 128	 accuracy: 0.959127 recall: 0.950204 f1 score: 0.951769	 accuracy: 0.981338 recall: 0.984391 f1 score: 0.978325
1024 - 256	 accuracy: 0.959658 recall: 0.950837 f1 score: 0.952404	 accuracy: 0.981997 recall: 0.984508 f1 score: 0.979068

2048 - 512	 accuracy: 0.960464 recall: 0.952046 f1 score: 0.953254	 accuracy: 0.981418 recall: 0.98457 f1 score: 0.978364
4096 - 1024	 accuracy: 0.959657 recall: 0.952075 f1 score: 0.952404	 accuracy: 0.979433 recall: 0.983746 f1 score: 0.976126

predict

- 테스트용으로 남겨둔 reading 음성 163개로 진행
- 모두 사람 목소리 데이터
- 상기 학습시킨 RF model로 테스트

```
X 개수:163, y 개수:163
                path
cate2
original_clean   163
정답률:  0.3128834355828221
```

y_pred	Non	voice	answer
Non_voice	112	112	112
voice	51	51	51

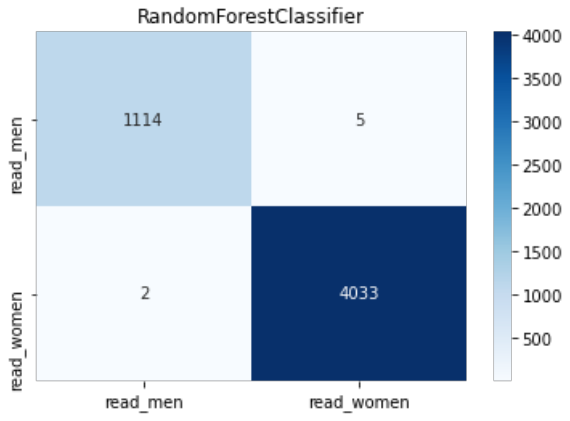
```
X 개수:163, y 개수:163
                path
cate2
original_clean   163
정답률:  0.31901840490797545
```

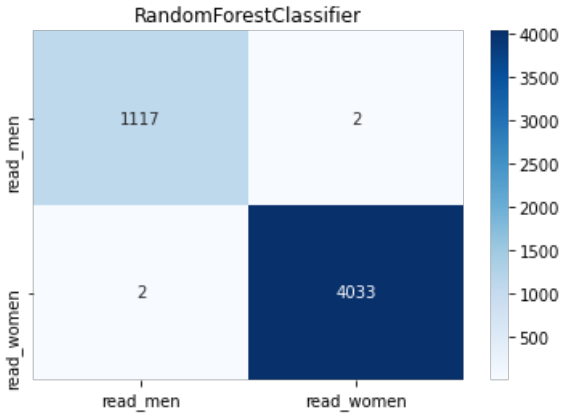
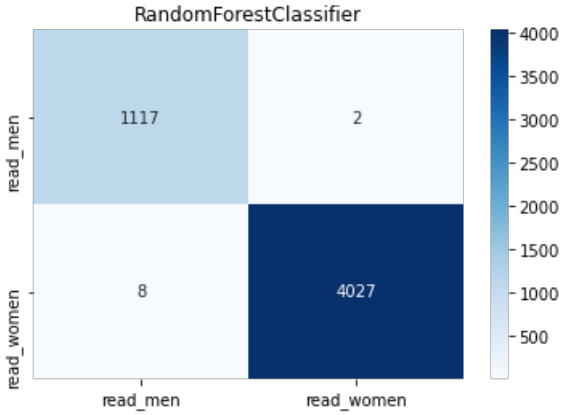
y_pred	Non	voice	answer
Non_voice	111	111	111
voice	52	52	52

M2 화자의 성별이 무엇인가?

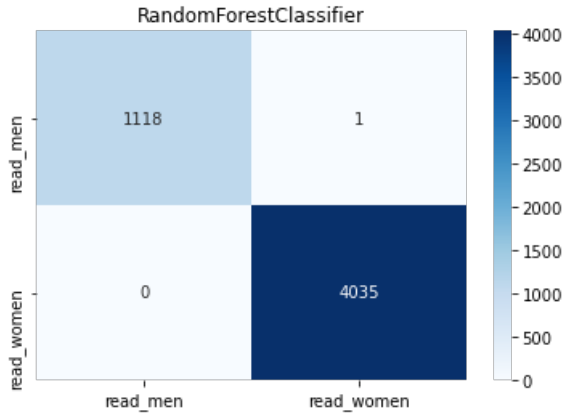
Train - test

- Train셋
  - 총 5.4만개
  - 여자 4만, 남자 1.4만
- test셋 - 파라미터별 결과 확인 → 모두 99%를 넘어서 Test 값만으로는 어떤 파라미터든 상관없어 보입니다.

n_fft - hop_length	RF
512 - 128	 accuracy: 0.998642 recall: 0.997518 f1 score: 0.998

1024 - 256	 accuracy: 0.999224 recall: 0.998859 f1 score: 0.998859
2048 - 512	 accuracy: 0.99806 recall: 0.998115 f1 score: 0.997152

4096 - 1024




```
accuracy: 0.998836
recall: 0.998934
f1 score: 0.99829
```

predict

- 테스트용으로 남겨둔 reading 음성 163개로 진행
- 직접 듣고 남자-여자 라벨링 진행

	answer
path	
../../sounds/human_voice/original_clean/c_001.mp3	read_women
../../sounds/human_voice/original_clean/c_002.mp3	read_men
../../sounds/human_voice/original_clean/c_003.mp3	read_women
../../sounds/human_voice/original_clean/c_004.mp3	read_women
../../sounds/human_voice/original_clean/c_005.mp3	read_men
...	...
../../sounds/human_voice/original_clean/c_159.mp3	read_men
../../sounds/human_voice/original_clean/c_160.mp3	read_men
../../sounds/human_voice/original_clean/c_161.mp3	read_women
../../sounds/human_voice/original_clean/c_162.mp3	read_women
../../sounds/human_voice/original_clean/c_163.mp3	read_women

163 rows × 1 columns

- 상기 학습시킨 RF model로 테스트
  - n_fft가 512 일 때 63.2%로 그나마 가장 성적이 좋습니다.

```
X 개수:163, y 개수:163
                path
cate2
original_clean   163
정답률:  0.6319018404907976
```

		y_pred
answer	TF	
read_men	False	37
	True	54
read_women	False	23
	True	49

```
X 개수:163, y 개수:163
                path
cate2
original_clean   163
정답률:  0.5828220858895705
```

		y_pred
answer	TF	
read_men	False	47
	True	44
read_women	False	21
	True	51

```
X 개수:163, y 개수:163
                path
cate2
original_clean   163
정답률:  0.5644171779141104
```

		y_pred
answer	TF	
read_men	False	49
	True	42
read_women	False	22
	True	50

신규 음원으로 predict 테스트

- 결과를 영상에 자막으로 입혀 보여드릴 예정입니다.
- 위 predict결과에서 보시다시피 성능이 그리 좋지 않기에.. voice이면 gender 분류하는 단계로 가기보다, 각 모델의 모든 결과값을 보는 것이 좋다고 판단했습니다.
- 충무로의 경우 대부분 사람 목소리만 나오기 때문에, gender 분류에 적합하다고 보여 먼저 테스트 진행했습니다.

M2 - gender 분류

시도1 : 전체 데이터로 학습시킨 모델 → 거의 랜덤한 결과를 보여주는 수준

- 절망하면서 동시에 든 생각
- 위에 163개로 미니 predict해보았을때, n_fft별로 정확도가 꽤 차이가 났는데
    - 성별 구분에 있어서는 512가 최적인게 아닐까?
- train셋도 전체가 아니라 512만 가져와서 모델 재학습 해보자

시도2: nfft 512로 학습시킨 모델

- train - test의 파라미터를 통일시켜 학습 및 predict
- 그 중 정확도가 제일 높은 512로 충무로 predict 테스트를 진행

```
512 RF
X 개수:163, y 개수:163
                path
cate2
original_clean   163
정답률:  0.656441717791411
                y_pred
answer     TF
read_men   False      42
           True       49
read_women False      14
           True       58
```

```
1024 RF
X 개수:163, y 개수:163
                path
cate2
original_clean   163
정답률:  0.6319018404907976
                y_pred
answer     TF
read_men   False      38
           True       53
read_women False      22
           True       50
```

```
2048 RF
X 개수:163, y 개수:163
                path
cate2
original_clean   163
정답률:  0.6441717791411042
                y_pred
answer     TF
read_men   False      37
           True       54
read_women False      21
           True       51
```

- 참고: 4096 - 52.7%로 생략

충무로_512_gender.mp4

M1 - voice 분류

시도1 : 목소리 or not 이진분류 모델

- 위 predict에서 이미 30%의 정확도라서 걱정되는 상태로 해보았는데
  대체로 목소리가 아니라고 하는 결과물..

충무로_1m_4096_voice.mp4

- 혹시 배경음이 크게 들려서 그런가?
    - 학습데이터가 사람 목소리 vs 그 외(노래, 사람소리, 생활잡음, 자연)

시도2: 대분류를 라벨로 재학습

- voice or not 이진분류로 한 결과가 좋지 않아서 라벨을 뭉치지말고, 대분류로 세분화하여 상세 proba를 보고자함
- train-test 파라미터를 통일해서 테스트했을때, 그 중 정확도가 제일 높았던 1024로 충무로 predict 테스트를 진행

충무로_1024_voice.mp4hong_1024_voice.mp4

결론

n_fft 파라미터

- train-test에서는 파라미터별 성능 차이가 근소해 영향이 없어보였고,
  train에 모든 파라미터값을 다 넣고 하도 성능에 차이가 없었다
- 하지만 실제 데이터로 predict 해보니,
  predict 할 데이터의 파라미터와 train의 파라미터 값이 같을 때 성능이 제일 좋다.
- 사람 목소리 여부를 가를 때는 1024, 사람 목소리 중에서 남/여를 가를 때는 512가 가장 성능이 좋았다.
- 이는 아마도 음원의 특성과 구분하고자하는 라벨에 따라 조금씩 상이할 것으로 보인다.

모델링 프로세스

- 사람인지 아닌지 구분하는 것이 아직 정확도가 많이 좋지 않아서, 목소리일 때 성별 판단으로 넘어가는 프로세스가 적용되면 최종 유저입장에서는 매우 부정확한 결과로 보일 것 같다.
    - 현재 수준은 목소리와 노래, 기타 다른 소리의 구성비율로 이 영상의 분위기 (인터뷰, 다큐멘터리, 음악방송..?) 나 화자의 성별 비중 정도 유추해볼 수는 있는 수준인 것 같다.