소리 - 성별 분류 모델 병합 로직

음원 분류 머신러닝 영상 테스트 에서 보여드린 M1 (소리 분류), M2 (성별 분류)의 결과값을 효율적으로 보여주기 위한 로직 설계 과정을 기록합니다.

모델 학습 스펙

- M1 (소리 분류): RandomForest - nfft 1024, hop_length 256
- M2 (성별 분류): RandomForest - nfft 512, hop_length 128

프로세스

freqeuncy feature 데이터셋 만들기

- 기존 테스트 영상 : 충무로, 홍사운드
- 추가 영상
  - 예능: https://www.youtube.com/watch?v=VQSjZiU-UdU&t=385
  - vlog: https://www.youtube.com/watch?v=try6rRGsdrw&t=34s

M1, M2 에 각각 predict 진행

여기를 클릭하여 펼치기...

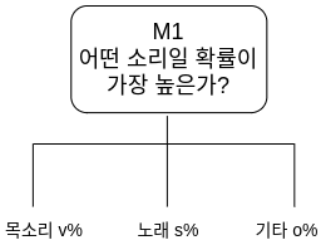
```
def proba_df(result_df, voice_nfft=1024, gender_nfft=512):
    voice_df = pd.DataFrame()
    for i in result_df[result_df['model']==1][result_df['nfft']
==voice_nfft].index:
        rf_test = pd.DataFrame(result_df['proba'].loc[i], columns=
['human', 'human_voice', 'life', 'nature', 'song'])
        rf_test['others'] = rf_test[['human', 'life', 'nature']].sum
(axis=1)
        rf_test['time'] = result_df['time'].loc[i]
        rf_test['pred_voice'] = result_df['y_pred'].loc[i]
        voice_df = pd.concat([voice_df, rf_test])
    voice_df.sort_values(['time'],inplace=True)

    gender_df = pd.DataFrame()
    for i in result_df[result_df['model']==0][result_df['nfft']
==gender_nfft].index:
        rf_test = pd.DataFrame(result_df['proba'].loc[i], columns=
['men', 'women'])
        rf_test['time'] = result_df['time'].loc[i]
        rf_test['pred_gender'] = result_df['y_pred'].loc[i]
        gender_df = pd.concat([gender_df, rf_test])
    gender_df.sort_values(['time'],inplace=True)

    merged_df = voice_df.merge(gender_df, on=['time'])
    return merged_df
```

- 각각 predict 후, time을 기준으로 concat
  - M1 결과값: human ~ pred_voice컬럼
  - M2 결과값: men, women, pred_gender 컬럼

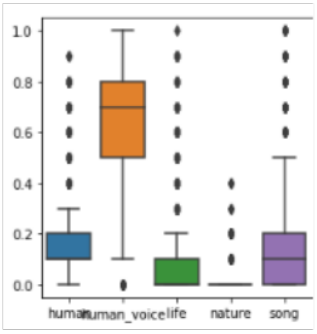
	human	human_voice	life	nature	song	others	time	pred_voice	men	women	pred_gender
0	0.2	0.7	0.1	0.0	0.0	0.3	0:00:00	human_voice	0.0	1.0	read_women
1	0.2	0.7	0.1	0.0	0.0	0.3	0:00:00.100000	human_voice	0.0	1.0	read_women
2	0.2	0.7	0.1	0.0	0.0	0.3	0:00:00.200000	human_voice	0.1	0.9	read_women
3	0.3	0.6	0.1	0.0	0.0	0.4	0:00:00.300000	human_voice	0.0	1.0	read_women
4	0.3	0.5	0.0	0.0	0.2	0.3	0:00:00.400000	human_voice	0.0	1.0	read_women
...	...	...	...	...	...	...	...	...	...	...	...
15725	0.3	0.1	0.3	0.0	0.3	0.6	0:26:12.500000	human	0.8	0.2	read_men
15726	0.3	0.1	0.3	0.0	0.3	0.6	0:26:12.600000	human	0.7	0.3	read_men
15727	0.3	0.0	0.4	0.0	0.3	0.7	0:26:12.700000	life	0.7	0.3	read_men
15728	0.4	0.0	0.3	0.0	0.3	0.7	0:26:12.800000	human	0.7	0.3	read_men
15729	0.4	0.0	0.3	0.0	0.3	0.7	0:26:12.900000	human	0.7	0.3	read_men

M1, M2 병합 로직 기획

1. 모델 predict



2. 영상 주요 음성 비중 확인

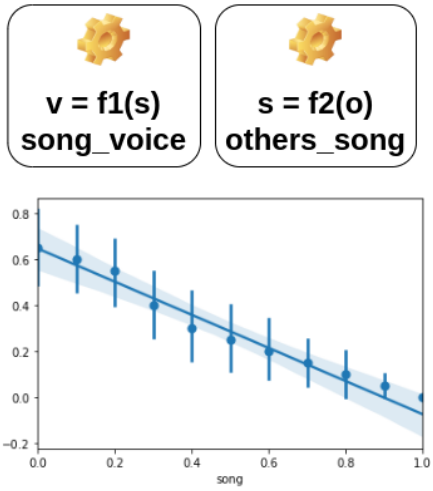


3. 목소리, 노래 Min baseline 설정

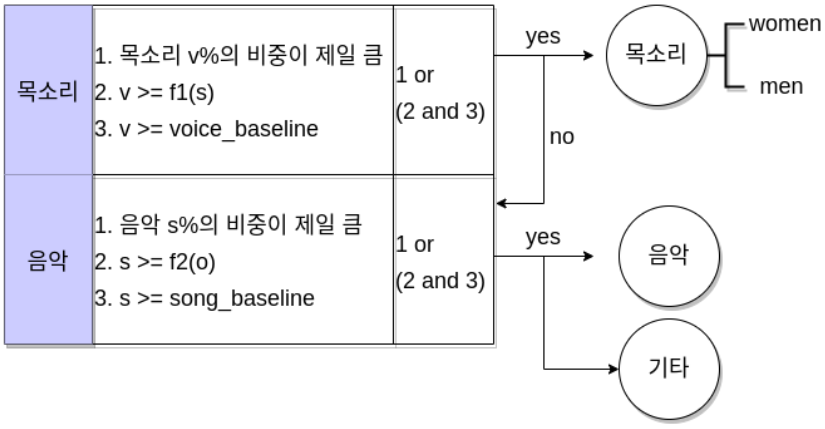
(백분율 기준)

구분	목소리	음악
voice_baseline	15%	50%
song_baseline	50%	25%

4. 영상 내 요소 비중 함수 생성

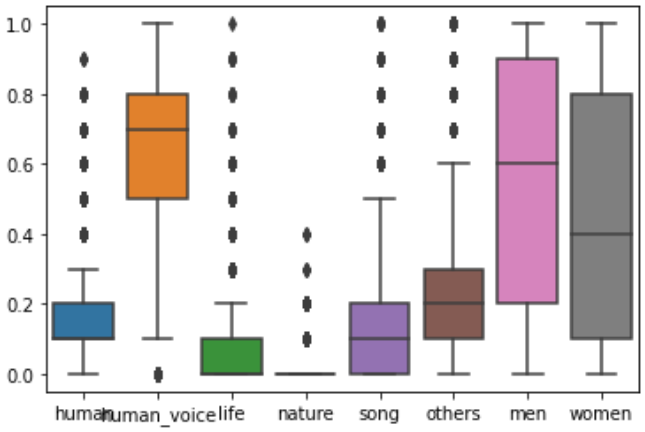


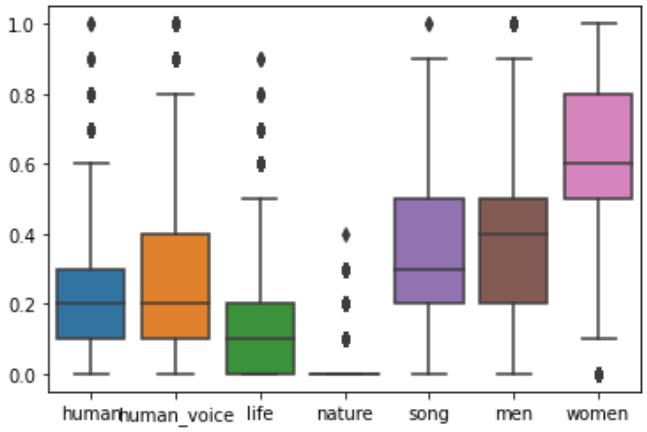
5. 로직 플로우



영상의 음성 비중에 따라 Voice와 Song 베이스라인 설정

- 추후 기획에 따라 유저가 직접 이 민감도를 조정할 수도 있도록 해도 좋을 듯
- 인터뷰, 대사가 위주인 영상 (ex: 충무로)
- 음성보다는 배경음 등 노래가 위주인 영상 (ex: 예능, 브이로그 등)





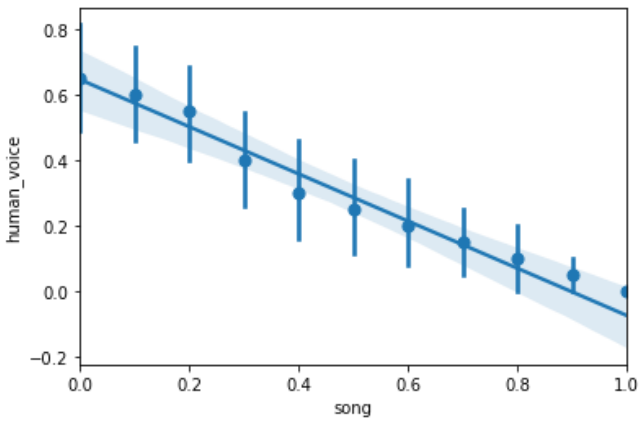
- 코드

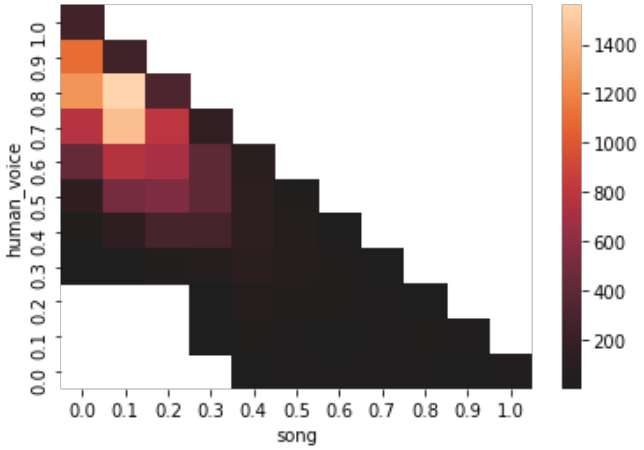
```
if np.argmax(merged_df.describe().loc['mean'][:5]) == 1: #human_voice
    print("Voice    .")
    voice_baseline = np.percentile(merged_df['human_voice'], 15)
    song_baseline = np.percentile(merged_df['song'], 50)
elif np.argmax(merged_df.describe().loc['mean'][:5]) == 4: #song
    print("Song    .")
    voice_baseline = np.percentile(merged_df['human_voice'], 50)
    song_baseline = np.percentile(merged_df['song'], 25)
```

영상 내 요소 상관관계 함수 생성

- 목소리와 노래 비중의 상관관계
  - pred_voice가 human_voice 인 데이터만 필터링
  - 코드

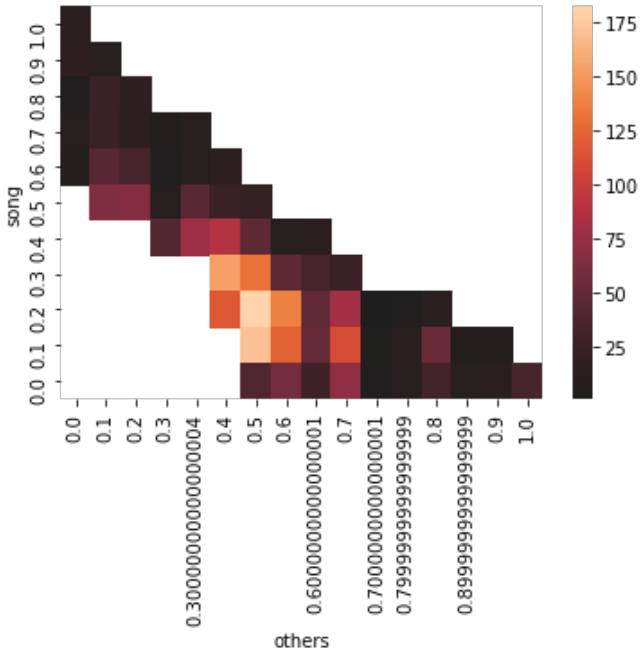
```
sns.regplot(y="human_voice", x="song", data=song_voice.
reset_index(),x_estimator=np.mean)
z = np.polyfit(song_voice.reset_index()['song'], song_voice.
reset_index()['human_voice'], 2)
song_voice_f = np.poly1d(z)
```

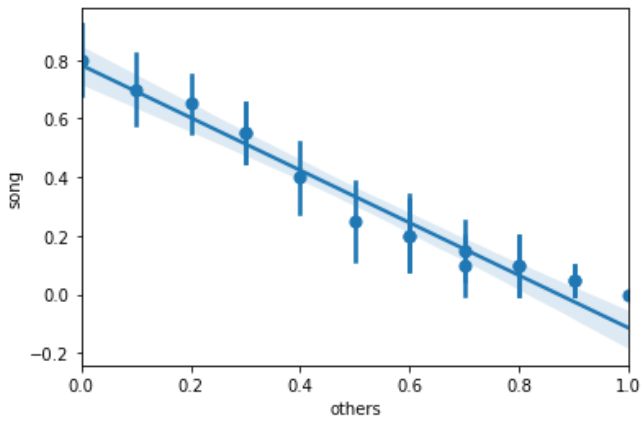




- 노래와 기타 소리 비중의 상관관계
    - pred_voice가 song과 others인 데이터 필터링
    - 코드

```
sns.regplot(x="others", y="song", data=song_others.
reset_index(),x_estimator=np.mean)
z = np.polyfit(song_others.reset_index()['others'],
song_others.reset_index()['song'], 1)
others_song_f = np.poly1d(z)
```





로직 플로우

- 위 프로세스 함수화

자막 로직 코드 전문

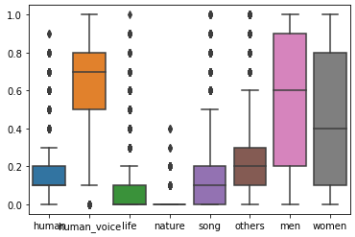
```
def create_subtitles(merged_df, song_voice_f, others_song_f):
    if np.argmax(merged_df.describe().loc['mean'][:5]) == 1:
#human_voice
        print("Voice    .")
        voice_baseline = np.percentile(merged_df['human_voice'], 15)
        song_baseline = np.percentile(merged_df['song'], 50)
    elif np.argmax(merged_df.describe().loc['mean'][:5]) == 4: #song
        print("Song    .")
        voice_baseline = np.percentile(merged_df['human_voice'], 50)
        song_baseline = np.percentile(merged_df['song'], 25)
    print("song : ",song_baseline, "/  : ", voice_baseline)
    st_ls = []
    for i in range(len(merged_df)):
        voice_pct = merged_df['human_voice'].iloc[i]
        song_pct = merged_df['song'].iloc[i]
        others_pct = merged_df['others'].iloc[i]
        pred_voice = merged_df['pred_voice'].iloc[i]
        if pred_voice=='human_voice' or ((voice_pct >= song_voice_f
(song_pct)) and (voice_pct>=voice_baseline)):
            pred_gender = merged_df['pred_gender'].iloc[i]
            if pred_gender == 'read_women':
                st_ls.append(f"Women  -- voice: {voice_pct*100}% /
song: {song_pct*100}%")
            else:
                st_ls.append(f"Men  -- voice: {voice_pct*100}% /
song: {song_pct*100}%")
        elif pred_voice=='song' or (song_pct>=song_baseline and
(song_pct >= others_song_f(others_pct))):
            st_ls.append(f"Song  -- voice: {voice_pct*100}% / song:
{song_pct*100}% / others: {round((1-voice_pct-song_pct)*100, 1)}%")
        else:
            st_ls.append(f"Others  -- voice: {voice_pct*100}% / song:
{song_pct*100}% / others: {round((1-voice_pct-song_pct)*100, 1)}%")
    return st_ls
```

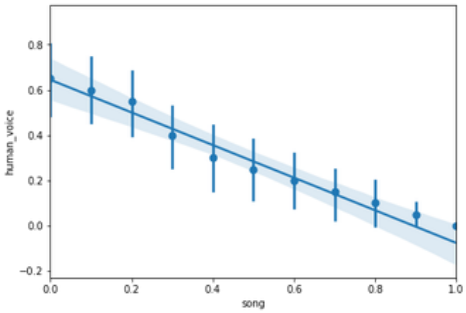
영상 테스트

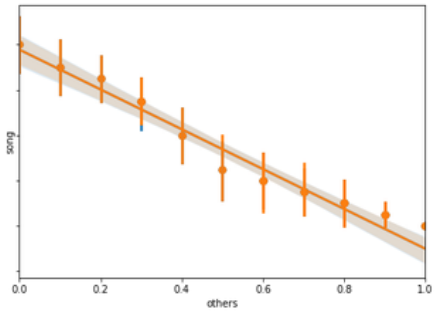
충무로

```
Voice가 주요 요소인 동영상 입니다.
song 베이스라인:  0.1 / 목소리 베이스라인:  0.4
```

	start_time	end_time	seq_num	srt_time	subtitle
0	00:00:00.000000	00:00:00.100000	1	00:00:00,000 --> 00:00:00,100	Women -- voice: 70.0% / song: 0.0%
1	00:00:00.100000	00:00:00.200000	2	00:00:00,100 --> 00:00:00,200	Women -- voice: 70.0% / song: 0.0%
2	00:00:00.200000	00:00:00.300000	3	00:00:00,200 --> 00:00:00,300	Women -- voice: 70.0% / song: 0.0%
3	00:00:00.300000	00:00:00.400000	4	00:00:00,300 --> 00:00:00,400	Women -- voice: 60.0% / song: 0.0%
4	00:00:00.400000	00:00:00.500000	5	00:00:00,400 --> 00:00:00,500	Women -- voice: 50.0% / song: 20.0%
...	...	...	...	...	...
15724	00:26:12.400000	00:26:12.500000	15725	00:26:12,400 --> 00:26:12,500	Song -- voice: 10.0% / song: 40.0% / others: ...
15725	00:26:12.500000	00:26:12.600000	15726	00:26:12,500 --> 00:26:12,600	Song -- voice: 10.0% / song: 30.0% / others: ...
15726	00:26:12.600000	00:26:12.700000	15727	00:26:12,600 --> 00:26:12,700	Song -- voice: 10.0% / song: 30.0% / others: ...
15727	00:26:12.700000	00:26:12.800000	15728	00:26:12,700 --> 00:26:12,800	Song -- voice: 0.0% / song: 30.0% / others: 7...
15728	00:26:12.800000	00:26:12.900000	15729	00:26:12,800 --> 00:26:12,900	Song -- voice: 0.0% / song: 30.0% / others: 7...







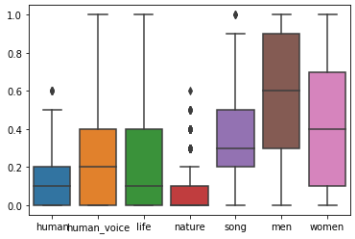
충무로_1m_1024_voice_1s.mp4

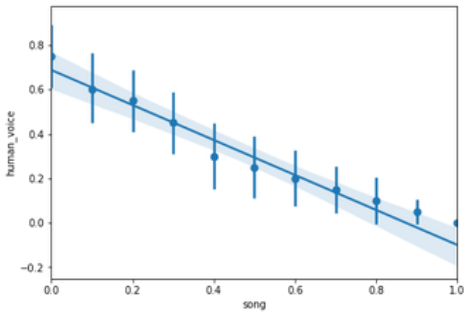
- 배경음이 있는 구간이 많아, 목소리와 음악 혼동 잦음 (배경음이 없으면 성별 인식도 더 잘 됨)
- 목소리에 음악을 합성시켜 학습을 시켜봐야하나..

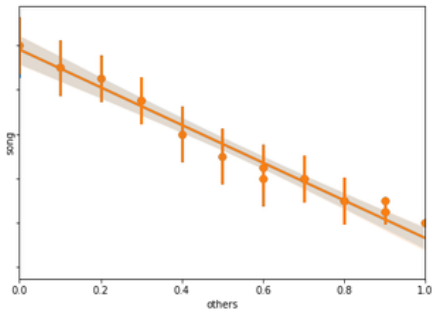
유학 준비 Vlog

```
Song이 주요 요소인 동영상 입니다.
song 베이스라인:  0.2 / 목소리 베이스라인:  0.2
```

	start_time	end_time	seq_num	srt_time	subtitle	srt
0	00:00:00.000000	00:00:00.100000	1	00:00:00,000 --> 00:00:00,100	Men -- voice: 50.0% / song: 30.0%	1\n00:00:00,000 --> 00:00:00,100\nMen -- voic...
1	00:00:00.100000	00:00:00.200000	2	00:00:00,100 --> 00:00:00,200	Men -- voice: 40.0% / song: 40.0%	2\n00:00:00,100 --> 00:00:00,200\nMen -- voic...
2	00:00:00.200000	00:00:00.300000	3	00:00:00,200 --> 00:00:00,300	Men -- voice: 60.0% / song: 30.0%	3\n00:00:00,200 --> 00:00:00,300\nMen -- voic...
3	00:00:00.300000	00:00:00.400000	4	00:00:00,300 --> 00:00:00,400	Men -- voice: 50.0% / song: 50.0%	4\n00:00:00,300 --> 00:00:00,400\nMen -- voic...
4	00:00:00.400000	00:00:00.500000	5	00:00:00,400 --> 00:00:00,500	Men -- voice: 60.0% / song: 40.0%	5\n00:00:00,400 --> 00:00:00,500\nMen -- voic...
...	...	...	...	...	...	...
2414	00:04:01.400000	00:04:01.500000	2415	00:04:01,400 --> 00:04:01,500	Song -- voice: 10.0% / song: 50.0% / others: ...	2415\n00:04:01,400 --> 00:04:01,500\nSong -- ...
2415	00:04:01.500000	00:04:01.600000	2416	00:04:01,500 --> 00:04:01,600	Song -- voice: 20.0% / song: 50.0% / others: ...	2416\n00:04:01,500 --> 00:04:01,600\nSong -- ...
2416	00:04:01.600000	00:04:01.700000	2417	00:04:01,600 --> 00:04:01,700	Others -- voice: 20.0% / song: 30.0% / others...	2417\n00:04:01,600 --> 00:04:01,700\nOthers -...
2417	00:04:01.700000	00:04:01.800000	2418	00:04:01,700 --> 00:04:01,800	Song -- voice: 30.0% / song: 40.0% / others: ...	2418\n00:04:01,700 --> 00:04:01,800\nSong -- ...
2418	00:04:01.800000	00:04:01.900000	2419	00:04:01,800 --> 00:04:01,900	Others -- voice: 30.0% / song: 20.0% / others...	2419\n00:04:01,800 --> 00:04:01,900\nOthers -...







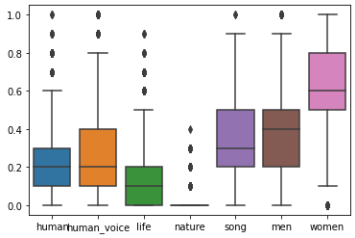
vlog_voice_0303.mp4

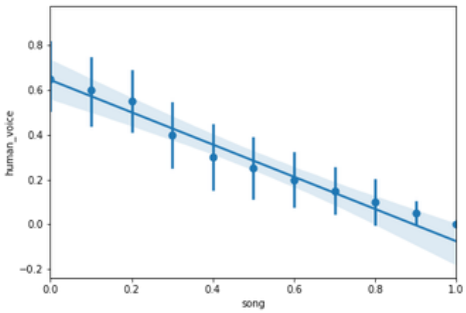
- 초반 배경음이 있는 구간을 남자 목소리라고 인식하는 오류
- 배경음이 없는 구간에서는 여자 목소리를 비교적 잘 잡으며,
  브이로그 특성상 목소리 없이, 노래 볼륨 낮게 화면만 나오는 경우엔 others로 잘 처리됨

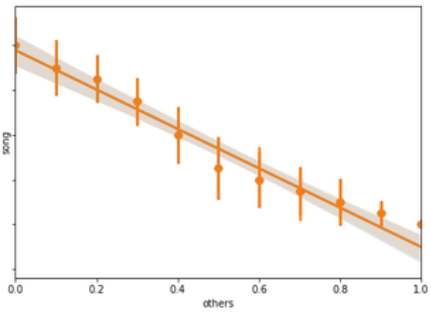
예능 산꾼도시여자들

Song이 주요 요소인 동영상 입니다.
song 베이스라인: 0.2 / 목소리 베이스라인: 0.2

	start_time	end_time	seq_num	srt_time	subtitle	srt
0	00:00:00.000000	00:00:00.100000	1	00:00:00,000 --> 00:00:00,100	Others -- voice: 40.0% / song: 10.0% / others...	1\n00:00:00,000 --> 00:00:00,100\nOthers -- v...
1	00:00:00.100000	00:00:00.200000	2	00:00:00,100 --> 00:00:00,200	Others -- voice: 40.0% / song: 10.0% / others...	2\n00:00:00,100 --> 00:00:00,200\nOthers -- v...
2	00:00:00.200000	00:00:00.300000	3	00:00:00,200 --> 00:00:00,300	Others -- voice: 30.0% / song: 30.0% / others...	3\n00:00:00,200 --> 00:00:00,300\nOthers -- v...
3	00:00:00.300000	00:00:00.400000	4	00:00:00,300 --> 00:00:00,400	Others -- voice: 40.0% / song: 10.0% / others...	4\n00:00:00,300 --> 00:00:00,400\nOthers -- v...
4	00:00:00.400000	00:00:00.500000	5	00:00:00,400 --> 00:00:00,500	Others -- voice: 30.0% / song: 10.0% / others...	5\n00:00:00,400 --> 00:00:00,500\nOthers -- v...
...	...	...	...	...	...	...
9164	00:15:16.400000	00:15:16.500000	9165	00:15:16,400 --> 00:15:16,500	Others -- voice: 20.0% / song: 0.0% / others:...	9165\n00:15:16,400 --> 00:15:16,500\nOthers -...
9165	00:15:16.500000	00:15:16.600000	9166	00:15:16,500 --> 00:15:16,600	Others -- voice: 20.0% / song: 0.0% / others:...	9166\n00:15:16,500 --> 00:15:16,600\nOthers -...
9166	00:15:16.600000	00:15:16.700000	9167	00:15:16,600 --> 00:15:16,700	Others -- voice: 20.0% / song: 10.0% / others...	9167\n00:15:16,600 --> 00:15:16,700\nOthers -...
9167	00:15:16.700000	00:15:16.800000	9168	00:15:16,700 --> 00:15:16,800	Others -- voice: 30.0% / song: 20.0% / others...	9168\n00:15:16,700 --> 00:15:16,800\nOthers -...
9168	00:15:16.800000	00:15:16.900000	9169	00:15:16,800 --> 00:15:16,900	Others -- voice: 40.0% / song: 10.0% / others...	9169\n00:15:16,800 --> 00:15:16,900\nOthers -...







산꾼도시여자들_0303.mp4

- 출연자들이 노래방기계로 노래를 부르는 장면이 많아 극강의 난이도..
- 리얼리티 예능이라 목소리가 작아서 others로 처리된 게 많음
- 노래 부르는 장면은 대체로 song으로 인식되며, women으로 처리된 경우도 있음