연구과제명 : 화자 분리를 이용한 음성 요약

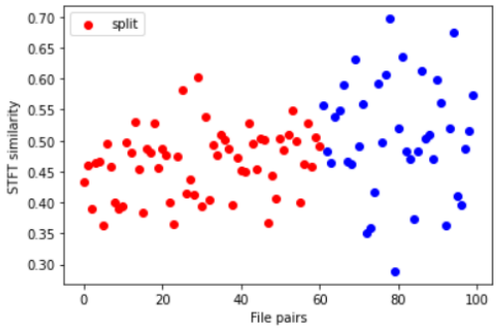
이번 주차에도 한 화자의 음성이 억지로 분리되는 상황을 방지하기 위하여 여러 시도를 해보았습니다.

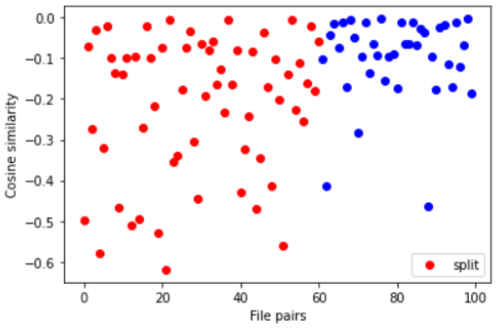
1. 화자 수 추정

   화자 수 추정 모델인 countnet을 사용하여 음성의 segment내에 총 몇 명의 화자가 있는지 먼저 구하고 모델의 결과에 따라 사용하는 화자 분리 모델의 channel을 결정하는 방식입니다. 위 모델의 정확도는 MAE값이 0.27로 나와있는데 실제로 해당 모델을 사용하여 시험적으로 적용해보니 오차율이 더 커서 환경 적인 문제가 있는지 확인해 보았고, 모델의 가중치를 제대로 불러오지 못하는 문제가 있다는 것을 확인하였습니다. 또한 해당 모델의 전이 학습이 가능한지 확인해 보았는데 아쉽게 trainning코드를 따로 제공해 주지 않았습니다. 그래서 일차적으로 해당 모델을 사용하는데 있어 발생하는 환경 문제들을 해결하고 MAE 0.27성능이 나오는 지 확인을 해봐야 할 것입니다.

2. 음성의 유사도 파악

   음성이 억지로 찢어지는 경우와 그렇지 않은 경우 어떠한 차이가 나는지 파악해 보기 위하여 KsponSpeech의 data일부를 화자 분리 모델에 넣어 위와 같은 문제 현상을 재현해 보았습니다. 총 100개의 음성을 분리하여 직접 들어봐서 해당 문제점이 발생하는 약60개의 찢어진 음성과 약 40개의 음성/노이즈로 분리된 음성에 대해서 분석을 해 보았습니다. 음성의 특징을 분석하는 방법으로는 STFT를 사용하였는데, 이 방법은 음성을 특정 길이의 구간으로 잘라서 각 구간별로 FFT를 적용하여 음성의 특징 벡터를 추출하는 방식입니다. 다음 추출된 두 벡터 사이의 코사인 유사도를 구하니 그 결과는 아래와 같았습니다. 또한 음성 자체의 배열 사이의 유사도를 구하니 다음과 같았습니다.





   결과적으로 한 음성이 2개로 찢어지는 형태이든 음성/노이즈로 찢어지는 형태이든 정량적인 방법으로 특징을 추출하는 것은 다른 방법이 필요해 보였습니다. 고로 저희 조의 이러한 이슈를 해결할 더 좋은 방법이 있는지 고민을 해 봐야 할 것입니다.

작성자	일자 2023-03-30 서명 신원철	확인자	일자 2023-03-30 서명 김유성