침수자
프로젝트

16기 신인섭, 천원준
17기 임청수, 홍예빈

Pipeline

침착맨 영상

화자 분리

Source 1

Source 2

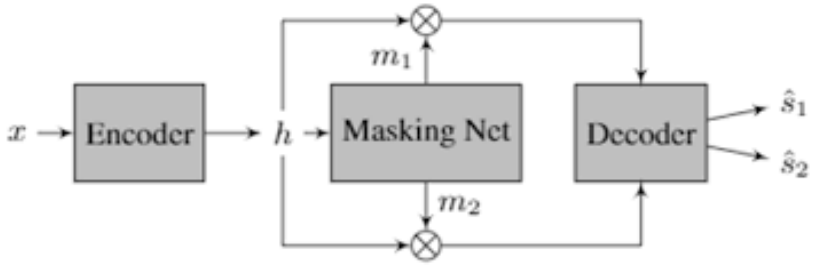
음성 인식

Text 1

Text 2

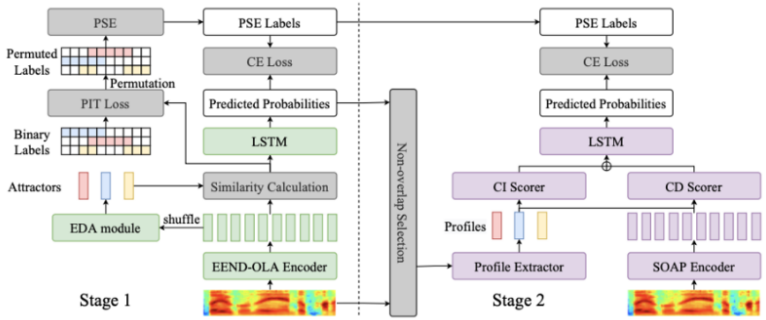
모델 선정 - 화자 분리(Speech Separation)

- SepFormer : Attention 기반 화자 분리



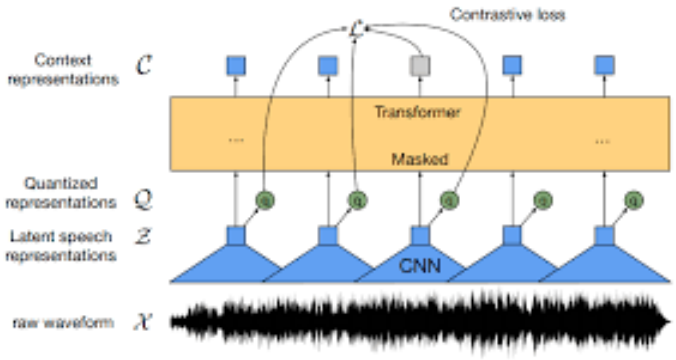
SepFormer 선택

- TOLD: 화자 분할 및 인식

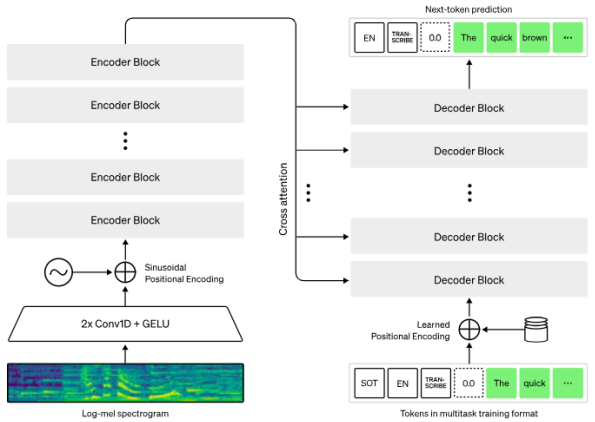


모델 선정 - 음성 인식(Speech Recognition)

- wav2vec 2.0



- Whisper



Whisper 선택

모델 학습 및 결과 - 화자 분리(SepFormer)

남 vs 남

침착맨

김풍

<

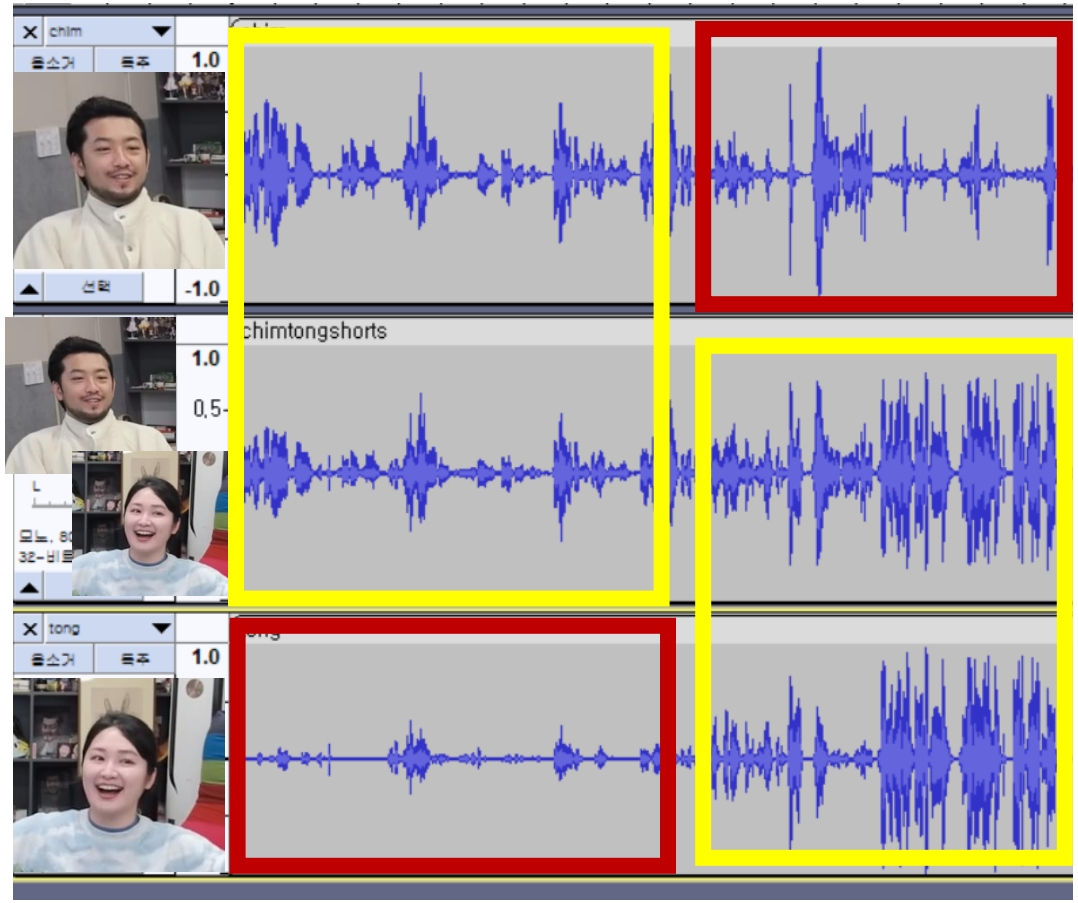
남 vs 여

침착맨

통닭천사

주파수와 음색의 차이로 인한 성능 차이

모델 학습 및 결과 - 음성 인식 전 데이터 처리



완벽한 분리 x, 노이즈 존재

Denosing process

모델 학습 및 결과 - 음성 인식(Whisper)

- finetuning X

침착맨

(0.0, 2.92) : ' 그 정도 빈도에 위치면'
(2.92, 5.1) : ' 슈.. 큰게 나요'
(10.6, 11.68) : ' 수고하셨어'
(11.68, 12.72) : ' 피키야 됐어'
(12.72, 13.44) : ' 해봐'
(13.44, 14.8) : ' 아 안녕하세요'
(14.8, 15.78) : ' 닭 닭 닭 닭'
(19.18, 21.38) : ' 네! 아니 뭐 4인 줄까요?'
(24.0, 26.0) : ' 아 유튜브에서 보셨을끼서'
(26.0, 36.0) : ' 아 지금 보인이때문에'
(36.0, 39.2) : ' 비빔들은 나 나오고'
(39.2, 41.2) : ' 에, 유튜브 나'
(41.2, 51.2) : ' 아, 혹시 누구지?'
(48.0, 50.0) : ' 아, 알겠습니다.'
(50.0, 52.0) : ' 아, 알겠습니다.'
(52.0, 54.0) : ' 아, 알겠습니다.'
(54.0, 56.0) : ' 아, 알겠습니다.'
(56.0, 58.0) : ' 아, 알겠습니다.'
(58.0, 60.0) : ' 아, 알겠습니다.'
(60.0, 62.0) : ' 아, 알겠습니다.'
(62.0, 64.0) : ' 아, 알겠습니다.'
(64.0, 66.0) : ' 아, 알겠습니다.'
(66.0, 68.0) : ' 아, 알겠습니다.'
(68.0, 70.0) : ' 아, 알겠습니다.'
(70.0, 72.0) : ' 아, 알겠습니다.'
(72.0, 74.0) : ' 아, 알겠습니다.'
(74.0, 76.0) : ' 아, 알겠습니다.'
(76.0, None) : ' 아, 알겠습니다. 아, 알겠습니다. 아, 알겠습니다. 아, 알겠습니다. 아, 알겠습니다. 아, 알겠습니다.'

통닭천사

(0.0, 2.0) : ' 아저씨가 나한테는 그...'
(2.0, 4.0) : ' 순줄기능...'
(4.0, 6.0) : ' 오마트는 어떻게 하자?'
(10.0, 12.0) : ' 어이, 아저씨가 나한테는...'
(12.0, 14.0) : ' 어디 가요?'
(16.0, 18.0) : ' 나 어디서 많이 본 것 같애!'
(18.0, 20.0) : ' 시비에 나오지 않았나?'
(20.0, 22.0) : ' 너 어디서 본 것 같은데?'
(22.0, 24.0) : ' 아저씨가 나한테는...'
(24.0, 26.0) : ' 아저씨가 나한테는...'
(26.0, 28.0) : ' 아저씨가 나한테는...'
(24.0, 25.78) : ' TV에 나오지 않았나?'
(25.78, 27.58) : ' 또 온수 본거 같은데?'
(30.04, 32.48) : ' 아 목소리도 많이 들어본거 같고'
(32.48, 33.72) : ' 어디 나왔지?'
(35.4, 36.54) : ' 아니...'
(36.8, 37.84) : ' TV...'
(41.02, 42.02) : ' 유튜브?'
(42.02, 43.36) : ' 유튜브 이름이 뭔데?'
(43.36, 44.92) : ' 내가 구독해 줄게요!'
(44.92, 46.02) : ' 진창맨이로'
(46.02, 48.5) : ' 나는 그 낚시하는 사람'
(48.5, 50.24) : ' 그게 좀 재밌었는데'
(51.52, 53.72) : ' 내가 자주 보는 거 있어'
(48.0, 50.0) : ' 그 사람은 그게 좀 생기판대'
(52.0, 54.0) : ' 내가 자주 보는 거 있어'
(54.0, 56.0) : ' 입질의 추억이라고'

모델 학습 및 결과 - 음성 인식(Whisper)

- 1st try : 침착맨 원본 박물관 데이터

침착맨

(0.0, 5.16) : ' 그 정도 빈도의 **위투면 큰 게** 나아요’
(10.56, 16.2) : ' 이거 많이 **컸어** 택시**가 됐어**요 해봐 안녕하세요 **떡** 탑소’
(19.2, 22.2) : ' **뭐 보인 줄**까요’
(24.0, 26.0) : ' 유튜브에서 **구새끼**께서’
(28.0, 30.0) : ' 아 지금 **보인인데**’
(32.0, 34.0) : ' **비비**는 안 나오고 있대’
(34.0, 36.0) : ' **예 힙을 놔**’
(48.0, 50.0) : ' **납시** 누구지?’
(48.0, 62.44) : ' **낙식**도 굳이 아예 없을 수 거의 뭐 **손손**도 먹고 오죠 이렇게 대화하는 **소강** 되잖아’
(62.44, 68.96) : ' 일단 거야 만약에 이제 대화하는 게 싫다 말이 끊기는 타임이 있어 **잘 가야 돼도**'

통닭천사

(0.0, 10.32) : ' 아저씨가 그냥 가게 안 돼 그 **순줄기** 있는 **오만큰한** 아프게 하자’
(10.32, 19.08) : ' 아이고 아이고 아이고 **쉬는** 거 어디 가요’
(19.08, 25.62) : ' 나 어디서 많이 본 것 같아 **시비**에 나오지 않았나’
(24.0, 25.78) : ' **시비**에 나 오지 않았나?’
(25.78, 27.7) : ' 너 오지 못 본 것 같은데'
(30.04, 33.76) : 목소리도 많이 들어본 것 같고 어디 나왔지'
(35.4, 37.88) : ' 아니 **시미**'
(41.04, 45.04) : ' 유튜브 **이벤트** 뭐인데 내가 구독해 줄게요'
(45.04, 46.02) : ' **진창맨**이로'
(46.02, 50.28) : ' 나는 그 낙시하는 사람 그게 좀 **생이던데**'
(48.0, 56.28) : ' 그래서 나면 그게 좀 **생이**던데 내가 자주 보는 거 있어 입질의 추억이라고'

추가 후처리 - 자막 생성

최종 모델 출력
script

자막 파일 변환

결과

한계

- 화자 분리 모델에서 1-2분이상의 input이면 성능도 좋지 못하고, 메모리의 한계
  -> 쇼츠 영상으로 실험하기로 결정
- 화자 분리 이후 음질 저하
- Whisper를 파인튜닝하는 과정에서 침착맨 데이터가 부족 + 메모리 한계

Thank you