침수자
프로젝트

16기 신인섭, 천원준
17기 임청수, 홍예빈

Pipeline

침착맨 영상

화자 분리

Source 1

Source 2

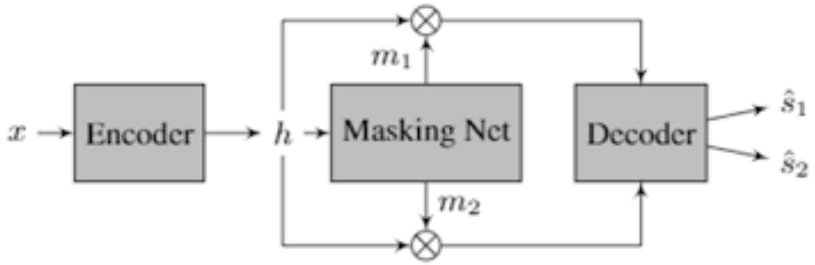
음성 인식

Text 1

Text 2

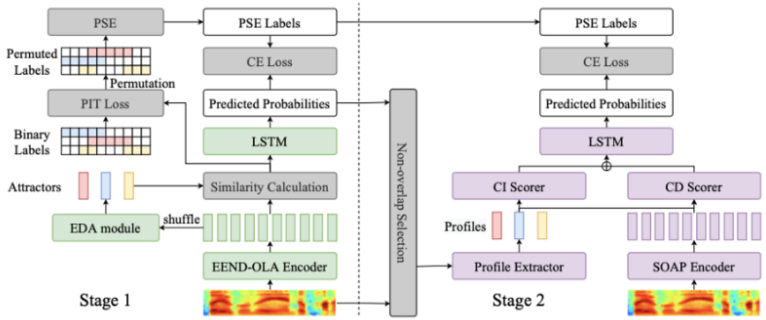
모델 선정 - 화자 분리(Speech Separation)

- SepFormer : Attention 기반 화자 분리



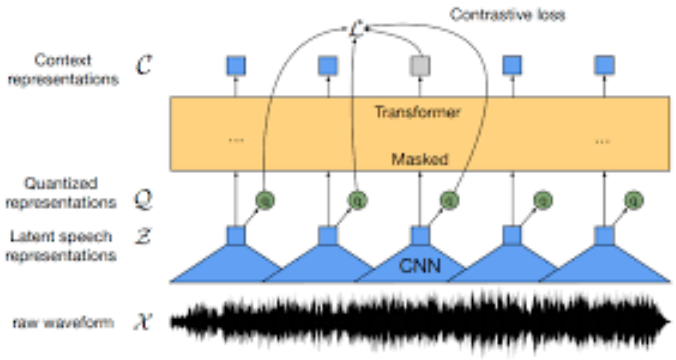
SepFormer 선택

- TOLD: 화자 분할 및 인식

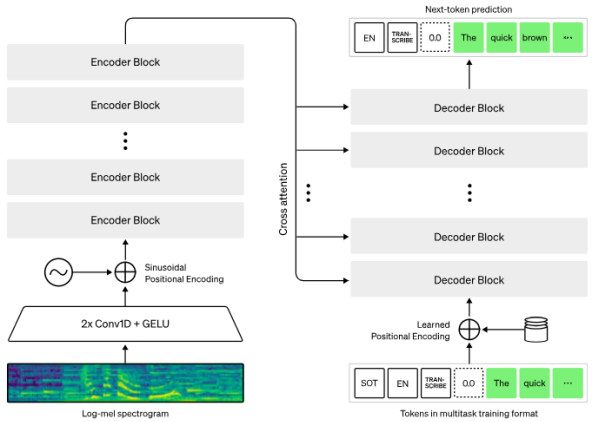


모델 선정 - 음성 인식(Speech Recognition)

- wav2vec 2.0



- Whisper



Whisper 선택

모델 학습 및 결과 - 화자 분리(SepFormer)

남 vs 남

침착맨

김풍

<

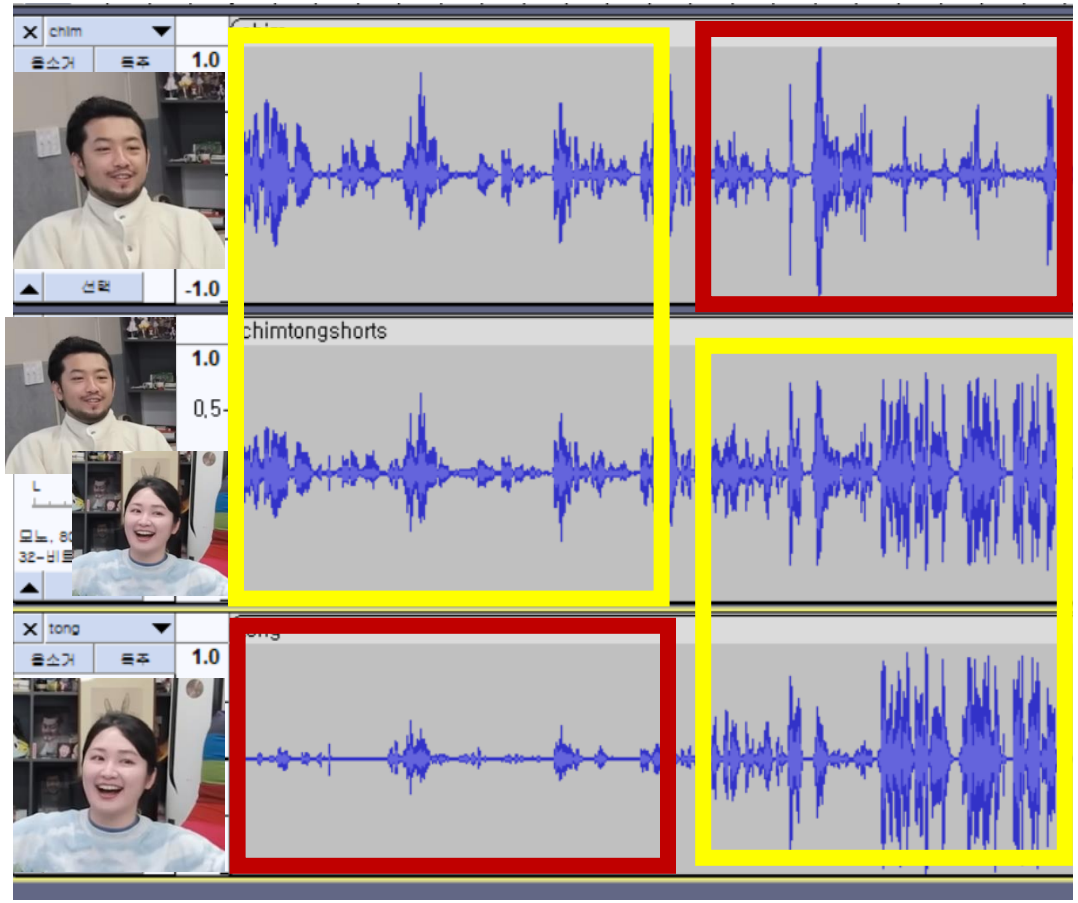
남 vs 여

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)

- 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'

모델 학습 및 결과 - 음성 인식(Whisper)

- 2nd try : 침착맨 원본 박물관 데이터 + 한국어 오픈 데이터

침착맨

(0.0, 5.16) : ' 그 정도 빈도의 위치면 택시타는 게 나아요’
(10.56, 16.2) : ‘ 이거 많이 탔어 택시가 탔어요 해봐 안녕하세요 딱 탑소’
(19.2, 22.2) : ' 네 뭐 부모님 댁가요’
(24.0, 26.0) : ' 유튜브에서 구새끼께서’
(28.0, 30.0) : ' 아 지금 부모님댁’
(32.0, 34.0) : ' 티비는 안 나오고 있대’
(34.0, 36.0) : ' 침착맨이라고’
(48.0, 50.0) : ' 납시 누구지?’
(48.0, 62.44) : ' 낙식도 굳이 아예 없을 수 거의 뭐 생선도 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 숨죽이 있는 오만큰한 어떻게 하자’
(10.32, 19.08) : ' 아이고 털보 친구 어디 가요’
(19.08, 25.62) : ' 나 어디서 많이 본 것 같아 티비에 나오지 않았나’
(24.0, 25.78) : ' 티비에 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 티비'
(41.04, 45.04) : ' 유튜브 이름이 뭐인데 내가 구독해 줄게요 '
(46.02, 50.28) : ' 나는 그 낙시하는 사람 그게 좀 생이던데'
(48.0, 56.28) : ' 그래서 나면 그게 좀 생이던데 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