AI 침투부 수호자 프로젝트

16기 신인섭, 천원준
17기 임청수, 홍여빈

Contents

01. 프로젝트 소개

01. 프로젝트 소개

프로젝트 목적

- ✓ 음성 인공지능을 활용하여 영상 자막을 자동 삽입해주는 프로젝트.
- ✓ 여러 명의 화자가 등장할 때 화자를 분할하고 겹치는 오디오를 분리하는 **화자인식**과 발화를 텍스트로 변환하는 **음성인식** 기술을 적용.

화자 인식 → 화자 : 쿠빅이

음성 인식 → 메시지 : 안녕

02. 프로젝트 파이프라인

02. 파이프라인

2021년 12월 31일 | 통닭천사와 떡볶이와 주먹밥 쿡방, 마카롱 먹방, 안산 배트...

조회수 135만회 • 2년 전

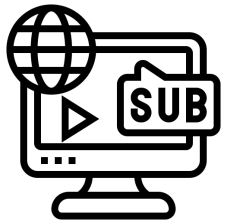
무편집 원본 영상

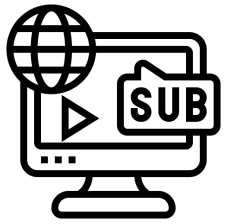
화자 인식

Source 1

Source 2

음성 인식





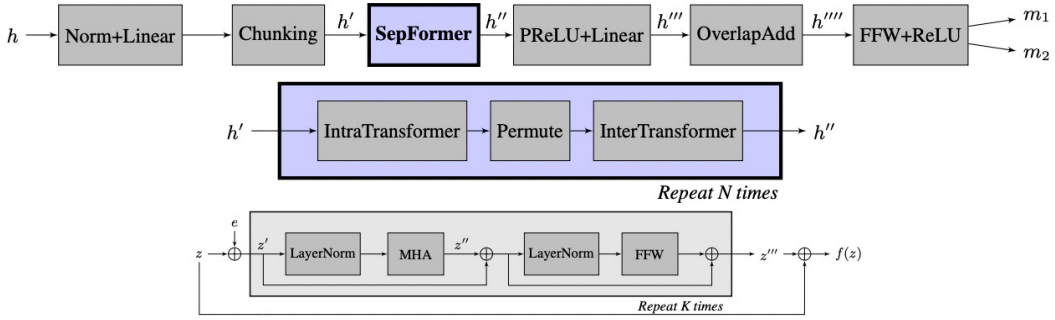
자막이 추가된 쇼츠 영상

03. 음성/화자 분리

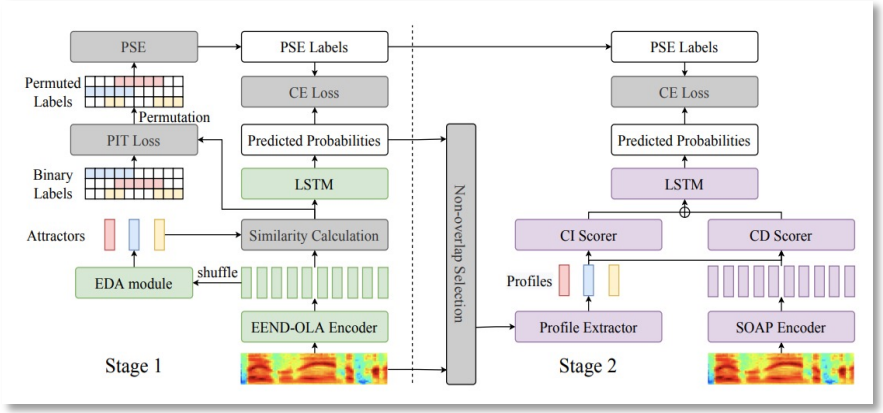
03. 음성/화자 분리

음성/화자 분리를 위한 모델

- ✓ SepFormer: Attention is All You Need in Speech Separation
- ✓ TOLD: A Novel Two-Stage Overlap-Aware Framework for Speaker Diarization



SepFormer 최종선택



TOLD

SepFormer : https://arxiv.org/abs/2010.13154
TOLD : https://arxiv.org/abs/2303.05397

03. 음성/화자 분리

프로젝트 대상 설정

- ✓ 주파수와 음색, 말투 등에서 특징을 추출하여 음성 분리를 수행하므로 남매 간 비교 시 성능이 더 우월함
- ✓ 본 프로젝트에서는 침착맨 남매가 대화하는 영상에 대해 자막 추출 진행

남 vs 남 비교

침착맨

김풍

남 vs 여 비교

침착맨

통닭천사

SepFormer : https://arxiv.org/abs/2010.13154
TOLD : https://arxiv.org/abs/2303.05397

KUBIG

03. 음성/화자 분리

화자분리 결과 해석

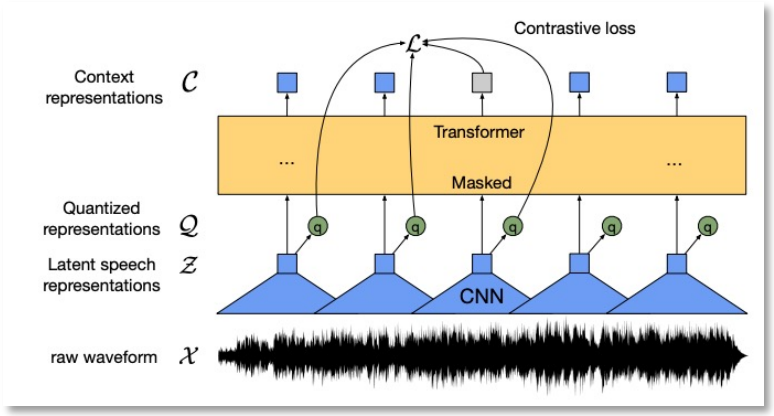
- ✓ SepFormer 모델을 통해 화자분리를 진행한 결과, 우측 사진처럼 화자별로 음성이 분리됨
- ✓ 하지만 다른 화자의 음성 크기가 줄어드는 방식으로 구현되므로 작은 크기의 음성은 디노이징 진행

04. 음성인식

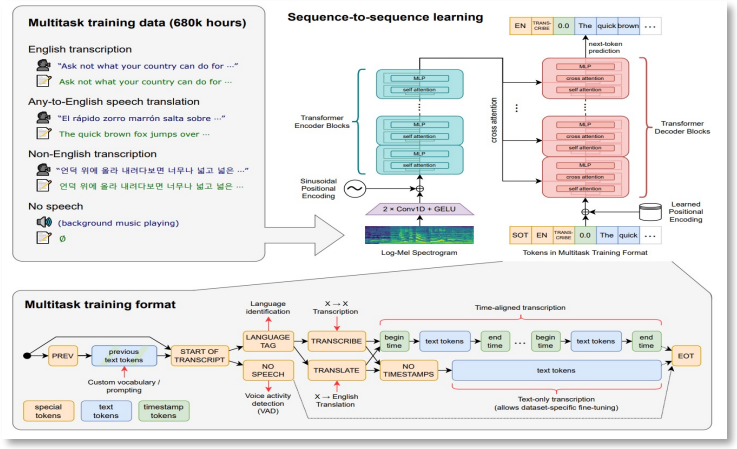
04. 음성인식

음성인식을 위한 모델

- ✓ wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations
- ✓ Whisper : Robust Speech Recognition via Large-Scale Weak Supervision



wav2vec



Whisper

최종선택

Wav2vec 2.0: https://arxiv.org/abs/2006.11477
TOLD : https://arxiv.org/abs/2303.05397

04. 음성인식

모델 선정 이유

- ✓ 1. Time stemp 추출 가능
- ✓ 2. 다국어 중 7번째로 많은 한국어 데이터로 학습하여 높은 성능 확보

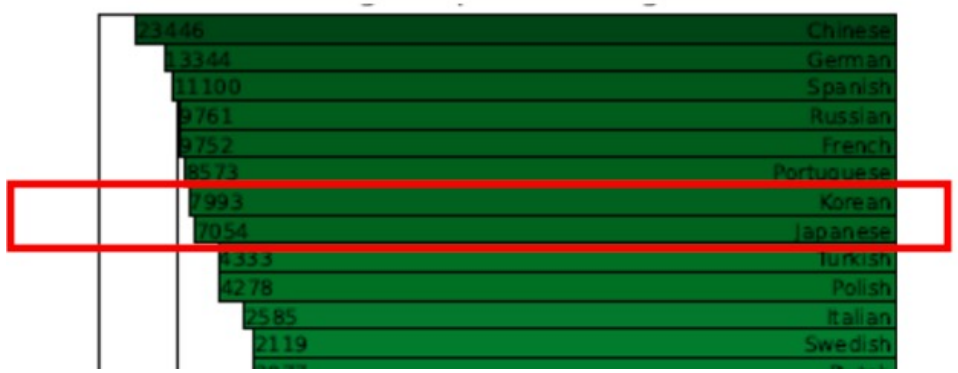
```
00:00:03,760 --> 00:00:10,500
아저씨 안녕하세요

3
00:00:10,500 --> 00:00:13,930
친구 어디 가요 가요

4
00:00:13,930 --> 00:00:18,600
나 어디서 많이 본 거 같아 이게 나오지 않았나

5
00:00:18,600 --> 00:00:19,800
또 어디서 본 거 같은데
```

자막 시점을 나타내는 Time Stemp



7번째로 많은 한국어 데이터

Wav2vec 2.0: https://arxiv.org/abs/2006.11477
TOLD : https://arxiv.org/abs/2303.05397

KUBIG

04. 음성인식

모델 학습하기 전

(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 줄까
(24.0, 26.0) : ' 아 유튜브에서 보셨을까'
(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.'

(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?'
(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'

04. 음성인식

침착맨 원본 박물관 데이터로 모델 학습 후

(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**'

(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'

KUBIG

04. 음성인식

자막 파일로 변환

- ✓ 1. Time stemp와 내용을 .srt 파일로 변환하여 자막 생성
- ✓ 유튜브 쇼츠에 자막 파일 삽입

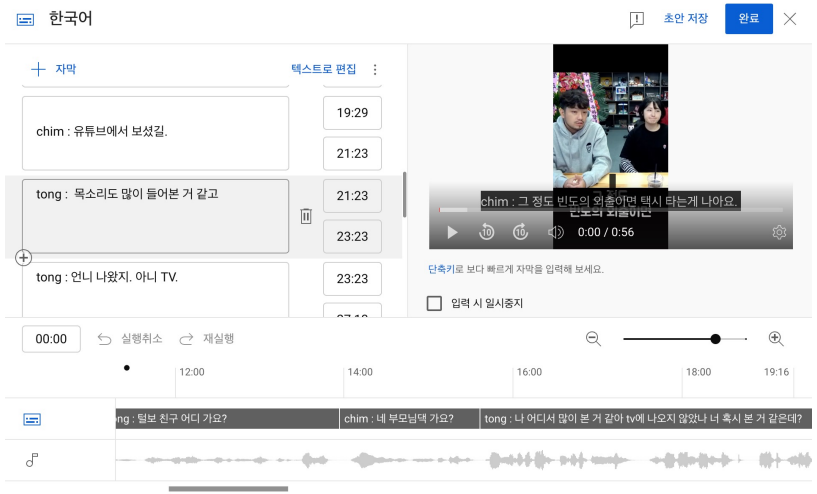
```
00:00:03,760 --> 00:00:10,500
아저씨 안녕하세요

3
00:00:10,500 --> 00:00:13,930
친구 어디 가요 가요

4
00:00:13,930 --> 00:00:18,600
나 어디서 많이 본 거 같아 이게 나오지 않았나

5
00:00:18,600 --> 00:00:19,800
또 어디서 본 거 같은데
```

자막 시점을 나타내는 Time Stemp



쇼츠에 자막 파일 삽입

Wav2vec 2.0: https://arxiv.org/abs/2006.11477
TOLD : https://arxiv.org/abs/2303.05397

KUBIG

05. 최종결과

05. 최종 결과

https://www.youtube.com/shorts/0eQuBnkHjRs

06. 결론 및 한계

06. 결론 및 한계

프로젝트 결론

- ✓ 화자 인식을 통해 영상 내에서 화자를 구분하여 자막 삽입
- ✓ 화자 인식과 음성 인식을 활용한 자막 추출로 영상 편집의 효율성 극대화

한계점

- ✓ 모델 메모리의 한계로 쇼츠 외의 긴 시간 영상 편집에 어려움 발생
- ✓ 화자분리 과정에서 음질 저하로 음성 인식 성능에 악영향
- ✓ 음성인식 모델 학습 시 대용량 데이터 확보가 어려워 기존 api에 비해 성능 차이 발생

Thank you