19 다중화자 탐지 구현과 STT 기술 결합

소속 정보컴퓨터공학부 분과 B 팀명 Untouchable 참여학생 변상윤, 문성필 지도교수 권준호

개요 및 목표

개요

최근 인공지능 기술의 발전으로 음성인식 기술인 STT (Speech-To-Text) 모델이 다양하게 개발되었다. STT란 사람이 말하는 음성 언어를 컴퓨터가 해석하여 텍스트로 변환하는 처리를 의미한다. 이는 회의록 작성, 유튜브 자막 생성, 상담 기록, 음성 명령어 처리, 청각 장애인들의 학습권 보장 등 다양한 분야에서 활용되고 있다.

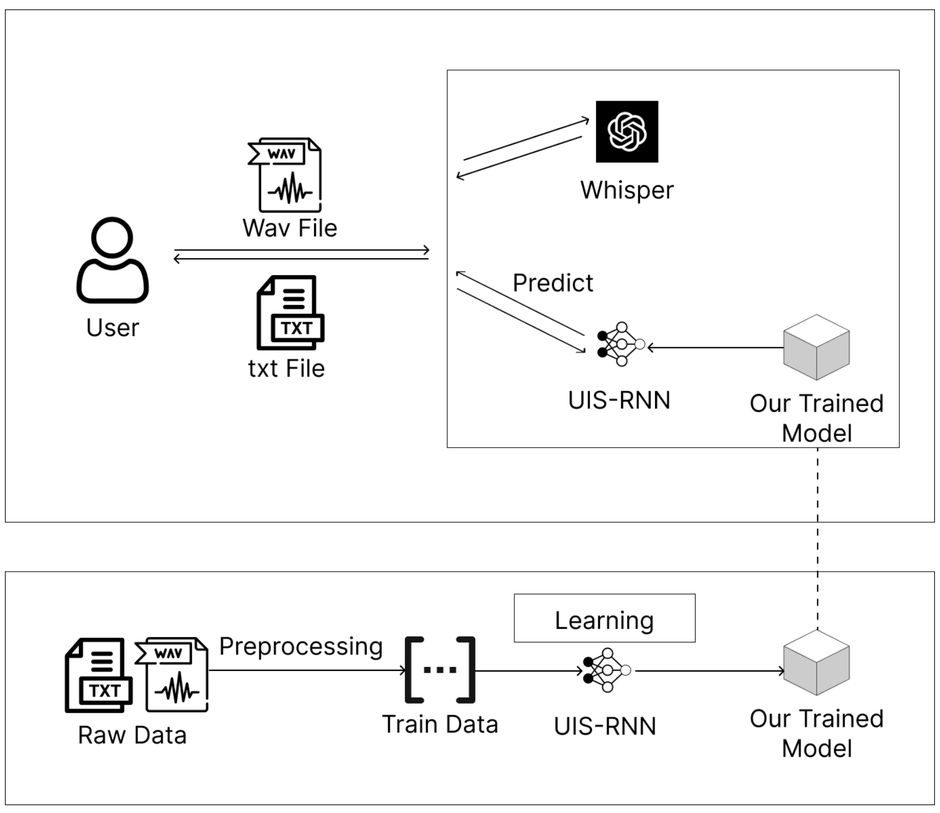
우리는 이러한 최신 STT를 이용하여 다중화자 기능을 추가함으로써 위의 효과와 더불어 추가적인 효용을 기대할 수 있다. 예를 들어 다중화자 인식에 대한 알고리즘 학습과 그에 대한 심화를 생각 해 볼 수 있다.

목표

다중화자 기술을 기존의 학습되어진 모델을 사용하지 않고 구현한다. 이 목표를 이루기 위해서 먼저 라벨링이 된 데이터셋을 받아 모델을 학습을 시키고, 이 학습된 모델을 이용해 화자 예측을 한다. 그리고 기존 STT 기술을 사용해 음성 파일을 대화록으로 바꾸어서 화자 정보와 대화 내용을 출력한다. 이 기술을 이용해 회의나 일상 대화의 자동 대화록 생성 기능을 하는 실용성 있는 애플리케이션을 개발한다. 이렇게 개발된 프로그램을 상용 프로그램과 비교를 해서 검증을 해본다.

상세 내용

구조도



다중화자 예측

```
Ground truth : 1 1 1 2 2 3
We predicted : 0 1 1 2 2 3
 0.0$ 5.5$ 0 $ 국민의 좌절감과 박탈감을 헤아리지 못했다. 검찰개혁에 혼신의 힘을 다하겠다.
 6.2$ 9.5$ 1 $ 이해찬 더불어민주당 대표가 오늘 기자간담회에서 한 말인데요.
 10.0$ 14.6$ 1 $ 조국 사태와 관련한 유감 표명이자 추선 의원들의 쇄신 요구에 대한 응답으로 보입니다.

 15.4$ 18.5$ 2 $ 한편으로 야당에 대해서도 한마디 했습니다.
 19.3$ 26.2$ 2 $ 정치 삼심해하면서 이런 야당 처음 본다. 꽉 막힌 패스트트랙 전국에 대해서 답답함도 표현했습니다.

 26.2$ 30.3$ 3 $ 네, 패스트트랙 충돌 사건을 수사하는 검찰은...
```

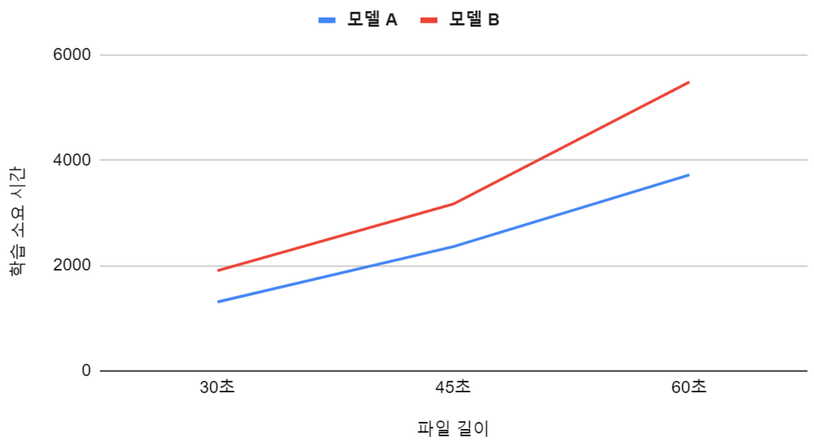
예측 결과

	모델 A			모델 B		
파일 길이	음성 파일 1	음성 파일 2	음성 파일 3	음성파일 1	음성 파일 2	음성 파일 3
30초	소요 시간:1479 정확도:0.145 문장별 정확도:0.600 (3/5)	소요 시간:1248 정확도:0.279 문장별 정확도:0.833 (5/6)	소요 시간:1203 정확도:0.197 문장별 정확도:0.800 (4/5)	소요 시간:2102 정확도:0.131 문장별 정확도: 0.400 (2/5)	소요 시간:1599 정확도:0.287 문장별 정확도:0.500 (3/6)	소요 시간:2009 정확도:0.115 문장별 정확도:0.400 (2/5)
45초	소요 시간:2606 정확도:0.153 문장별 정확도:0.500 (4/8)	소요 시간:2117 정확도:0.213 문장별 정확도:0.400 (4/10)	소요 시간:2354 정확도:0.204 문장별 정확도:0.625 (5/8)	소요 시간:3044 정확도:0.194 문장별 정확도: 0.500 (4/8)	소요 시간:2206 정확도:0.243 문장별 정확도: 0.400 (4/10)	소요 시간:4258 정확도:0.160 문장별 정확도:0.625 (5/8)
60초	소요 시간:3999 정확도:0.144 문장별 정확도:0.467 (7/15)	소요 시간:3258 정확도:0.203 문장별 정확도:0.385 (5/13)	소요 시간:3903 정확도:0.188 문장별 정확도:0.545 (6/11)	소요 시간: 5395 정확도: 0.115 문장별 정확도: 0.400 (6/15)	소요 시간: 3961 정확도: 0.166 문장별 정확도:0.308 (4/13)	소요 시간: 7096 정확도: 0.113 문장별 정확도:0.364 (4/11)

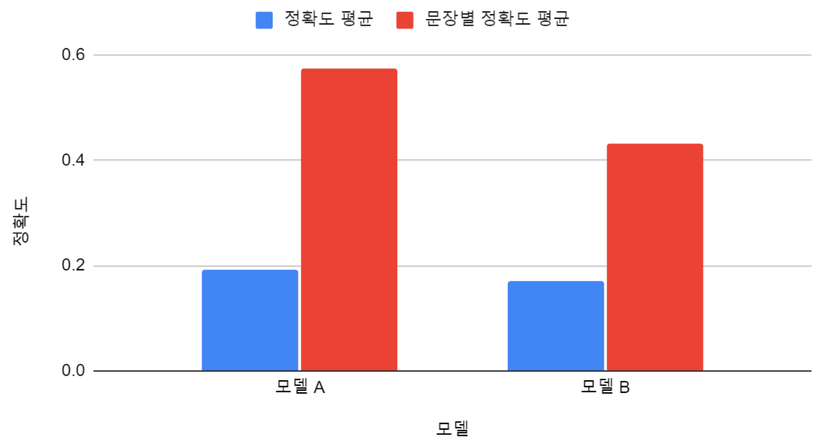
표 1. 모델A, 모델B의 결과 예측 소요시간, 정확도, 문장별 정확도에 따른 비교

결과

그래프 1. 파일 길이에 따른 예측 소요 시간



그래프 4. 정확도 평균과 문장별 정확도 평균



모델 A : 3개의 파일을 1000번 반복 학습 시킨 모델
모델 B : 75개의 파일을 50번 반복 학습 시킨 모델
정확도 : 화자 번호 리스트에서 요소별로 최적 적합(optimal matching)을 사용하여 비교한 정확도
문장별 정확도 : whisper의 타임스탬프에 맞춰 구간별 가장 많은 화자 번호를 대표 화자번호로 하여 문장별로 최적 적합을 사용하여 비교한 정확도

- **그래프 4를 보았을 때, 정확도 평균보다 문장별 정확도 평균이 현저하게 높다는 것을 알 수 있다. 이는 데이터 매핑 - 후처리가 중요하다는 것을 보여준다.**
- **그럼에도 불구하고 상용 애플리케이션에 비해 부족한 점이 많으므로 보완이 필요하다.**