STT 성능 향상을 위한 딥러닝 기반 발화 음성 분리학습

김보경, 양영준, 황용해, 김규헌
경희대학교
bellbpng@khu.ac.kr, romanfury9@gmail.com, hyh717@khu.ac.kr,
kyuheonkim@khu.ac.kr

Deep Learning-based Speech Voice Separation Training To Enhance STT Performance

Bokyoung Kim , Youngjun Yang, Yonghae Hwang, Kyuheon Kim
Kyunghee University

요 약

인공지능을 활용한 다양한 딥러닝 기술의 보급과 상용화로 오디오 음성 인식 분야에서도 음성 인식의 정확도를 높이기 위한 다양한 연구가 진행되고 있다. 최근 STT 를 위한 음성 인식 엔진은 딥러닝 기술을 기반으로 과거에 비해 높은 정확도를 보이고 있다. 하지만 예능 프로그램, 드라마, 스포츠 방송 등과 같이 비음성 신호와 음성 신호가 함께 녹음되는 오디오의 경우 음성 인식 정확도가 크게 낮아지는 문제가 발생한다. 이에 본 연구에서는 다양한 장르의 오디오를 음성과 음악을 분리하는 딥러닝 모델을 활용하여 음성 신호와 비음성 신호로 분리하는 방법을 제시하고, STT 결과를 분석하여 음성 인식의 정확도를 높이기 위한 연구 방향을 제시한다.

1. 서론

딥러닝 기술의 발전과 함께 인공지능 비서, 스마트 스피커 등 음성 인식 서비스의 정확도를 높이기 위한 연구가 활발히 진행되고 있다. 음성 인식 서비스의 경우 발화 음성을 텍스트로 변환하는 Speech-To-Text(STT) 기술을 필요로 한다. 최근 STT 를 위한 음성 인식 엔진은 딥러닝 기술을 기반으로 과거에 비해 높은 정확도를 보이고 있지만, 음성 데이터가 가지는 비음성 신호의 특성에 따라 음성 인식 정확도가 크게 낮아지는 문제가 발생한다.

본 연구에서는 Google Cloud Platform 의 Cloud Speech-to-Text API 를 사용하여 STT 결과를 비교해보았다. STT 는 음성모델 학습을 통한 음소의 인식과 언어모델 학습을 통한 문맥의 인식으로 동작한다. 여기서 음소를 인식하는 음성모델 학습과 문맥을 인식하는 언어모델 학습은 이미 충분한 학습이 이루어져서 비음성 신호의 비중이 매우 낮은 고품질의 오디오 데이터에서 STT 인식률은 95% 이상을 달성하였다[1]. 하지만 실생활에서 많이 접하는 예능 프로그램, 스포츠, 뉴스 등 다양한 장르의 오디오 특성은 음성 신호와 비음성 신호가 중첩되는 형태로 나타나며, 이러한 데이터의 STT 인식률은 크게는 50% 아래로 떨어지기도 한다[1]. 본 연구에서는 STT 성능 향상을 위해 오디오 데이터에서 발화 음성 신호를 효과적으로 분리할 수 있도록 딥러닝 기반의 오픈소스 모델을 사용하여 발화 음성 분리

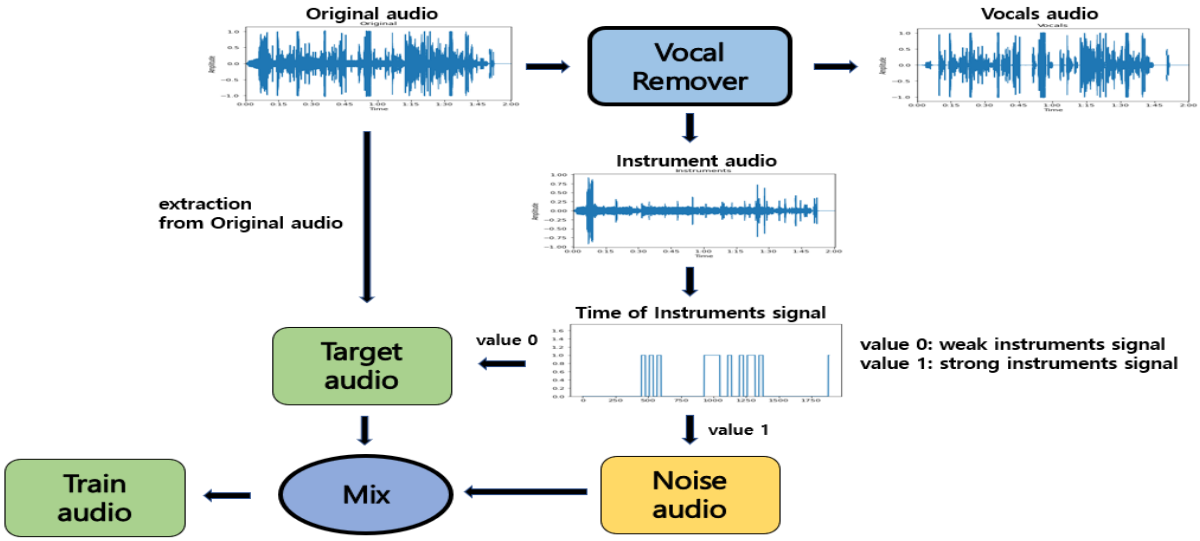

그림 1. 데이터 전처리 작업

학습을 진행하였다. 학습된 모델을 가지고 다양한 유형의 오디오 데이터를 전처리한 후 STT 결과를 도출했고 다른 데이터에 비해 평균적으로 더 많은 단어를 추출하는 것을 확인할 수 있었다.

2. 본론

(1) 비음성 오디오(Instruments) 추출

본 연구에서는 U-Net 구조의 DNN 을 사용하여 학습된 네트워크(Vocal Remover)를 이용했다[2]. Vocal Remover 는 입력 오디오로부터 음성 신호(Vocals)를 제거하여 비음성 신호(Instruments)를 추출하는 모델이다. 음성과 비음성 신호가 모두 존재하는 오디오에 학습된 네트워크를 사용하여 음성이 제거된 비음성 오디오를 생성하고, 입력 오디오와 비음성 오디오의 연산을 통해 음성 오디오를 생성한다.

(2) 데이터 생성과 모델 훈련

그림 1 은 전체적인 데이터 전처리 과정을 보여주고 있다. Vocal Remover 를 이용해 추출한 비음성 오디오(Instruments)의 신호의 세기를 적절한 임계치를 결정하여 비음성 신호가 약한 구간을 확인한다. 그리고 원본 오디오에서 비음성 신호가 약한 구간들을 추출하여 새로운 오디오 데이터를 생성한다. 이 구간의 오디오 데이터는 음성 신호가 비음성 신호에 비해 강하게 존재하는 구간이라고 추정할 수 있고 따라서 모델 훈련을 위한 정답(target) 데이터로 설정한다. 반대로 비음성 신호가 강한 구간들을 모아서 Noise 오디오를 생성하고 Voice 오디오와 섞는 작업으로 훈련(train) 데이터를 만든다. 준비된 데이터를 이용하여 Vocal Remover 의 훈련되지 않은 네트워크를 발화 음성 분리를 목적으로 학습한다.

(3) STT 결과 비교

본 연구의 STT 결과를 비교하기 위해 Google Cloud Platform 의 Cloud Speech-to-Text API 를 사용했으며, 그림 1 에서 제시하는 방법으로 얻은 데이터로 모델을 훈련시켜 얻은 Speech 오디오와 원본 오디오, Vocal Remover 로 얻어진 Vocals 오디오의 결과를 비교해보았다. 표 1 은 STT 결과를 평균 신뢰도와 단어 개수를 정량적 척도로 삼아 비교해본 결과이다. 음성 신호와 비음성 신호가 혼합된 오디오 데이터에서 비음성 신호를 최대한 제거하여 발화 음성 신호만을 분리하는 본 연구의 목적을 고려할 때, STT 로 추출된 단어 개수에서 다른 데이터들에 비해 좋은 성능을 보인다는 점은 효과적으로 발화 음성 분리 학습이 이루어졌다고 판단될 수 있다. 그러나 신뢰도가 다른 데이터에 비해 다소 떨어지는 현상은 추가적인 연구가 필요하다고 생각된다.

	Speech Separation		Original		Vocal Remover		
Content	Confidence	Words	Confidence	Words	Confidence	Words	유형
1	0.799518	638	0.835153	517	0.821507	631	드라마
2	0.798374	441	0.839062	424	0.764271	409	드라마
3	0.802707	151	0.751183	138	0.777239	137	드라마
4	0.81414	324	0.798689	331	0.8262	310	드라마
5	0.795391	887	0.831401	720	0.861748	749	드라마
6	0.779352	508	0.794423	476	0.856023	472	예능
7	0.754574	352	0.839071	345	0.774846	302	예능
8	0.789305	293	0.78287	278	0.771186	262	예능
9	0.823653	231	0.81037	229	0.758227	181	예능
10	0.842467	416	0.851178	380	0.849076	431	예능
11	0.88861	186	0.868692	185	0.88426	158	스포츠중계
12	0.881583	2641	0.86325	2597	0.873455	2538	스포츠중계
13	0.802037	226	0.855836	314	0.837523	259	정치/시사
14	0.811422	151	0.865993	200	0.736896	146	정치/시사

표 1. STT 결과 비교

3. 결론

본 연구에서 제안하는 비음성 신호가 강한 구간과 약한 구간을 파악하여 데이터를 전처리하고 발화 음성 분리를 목적으로 딥러닝 기반의 모델을 훈련시키는 방법은 STT 결과로 추출되는 단어 개수와 평균 신뢰도에 영향을 주는 것을 확인했다. 발화 음성 분리 모델로 추출된 오디오 데이터가 다른 데이터에 비해 평균적으로 더 많은 단어를 추출하는 것을 확인했고 이는 STT 음성 인식 엔진이 해당 데이터의 언어적 요소를 효과적으로 인식하고 있다고 판단된다.

입력되는 오디오 데이터의 유형 또한 STT 결과에 영향을 주는 것을 확인할 수 있었다. 드라마, 예능 프로그램 같이 다양한 비음성 신호가 혼합된 오디오 데이터의 경우 훈련된 모델을 이용해 음성 신호를 필터링한 데이터가 평균적으로 좋은 STT 성능을 보이지만 정치/시사 유형처럼 비음성 신호가 거의 없는 데이터의 경우 원본 데이터의 STT 성능이 더 좋은 것으로 확인됐다. 오히려 훈련된 모델로 필터링한 데이터의 음성 신호가 원본 데이터의 음성 신호에 비해 손상되는 것으로 판단된다. 더불어, 본 연구에서 제안한 방법으로 얻어진 데이터는 STT 결과의 평균 신뢰도가 다른 데이터에 비해 다소 떨어지는 모습을 확인할 수 있었고 이에 대한 추가적인 연구가 필요할 것으로 판단된다.

참 고 문 헌

[1] 최승주 and 김종배. (2017). 음성 인식 오픈 API 의 음성인식 정확도 비교 분석. 예술인문사회 융합 멀티미디어논문지, 7(8), 411-418.

[2] Andreas Jansson, Eric Humphrey “SINGING VOICE SEPARATION WITH DEEP U-NET CONVOLUTIONAL” (accessed Oct. 23-27, 2017).