STT 성능 향상을 위한 딥러닝 기반 음성 신호 필터

김보경, 이성배, 김규헌*
경희대학교

bellbpng@khu.ac.kr, rhee@khu.ac.kr kyuheonkim@khu.ac.kr

Deep Learning-based Filter for Speech Separation to Enhance STT Performance

Bokyoung Kim, Seongbae Lee, Kyuheon Kim*
Kyunghee Univ.

요 약

최근 딥러닝 기반 음성 인식 기술의 발달로 다양한 인공지능 서비스가 상용화되고 있으며, 그 중 음성 신호를 입력으로 하여 텍스트를 생성하는 STT(Speech-to-Text) 서비스가 대표적인 예시이다. 특히, 딥러닝 기반 STT 서비스는 높은 텍스트 생성 정확도를 바탕으로 다양한 OTT 서비스에서 활용되고 있지만, 예능 콘텐츠와 같이 배경음악과 효과음이 포함된 오디오의 경우에는 음성인식 정확도가 크게 저하되다는 제한사항이 있다. 이에 본 연구에서는 예능 콘텐츠의 배경음악과 효과음을 제거하여 딥러닝 기반 음성인식의 정확도를 높일 수 있는 음성 신호 필터를 제안하고, 실험을 통해 STT 성능 향상을 검증하고자 한다.

Ⅰ. 서 론

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 있으며, 일례로 Google Cloud Platform 의 Cloud Speech-to-Text API 를 통한 STT 인식률은 95% 이상을 나타낸다[1]. 이와 같이 높은 STT 정확도를 바탕으로 다양한 OTT 서비스에서 활용이 기대되고 있다.

일반적으로, STT 는 음성모델 학습을 통한 음소의 인식과 언어모델 학습을 통한 문맥 인식으로 동작한다. 여기서 음소를 인식하는 음성모델 학습과 문맥을 인식하는 언어모델 학습은 음성 신호 위주로 구성된 고품질의 오디오 데이터를 통해 진행되기 때문에 음성 위주의 신호에 대한 STT 정확도는 높게 나타난다.

그러나 오디오 데이터에 비음성 신호의 비중이 높아진다면 음성인식 정확도는 크게 저하될 수 있다. 일례로, 예능 콘텐츠와 같이 배경음악과 효과음 등의 비음성 신호가 포함된 오디오에서는 STT 인식률이 크게는 50% 아래로 나타나기도 한다[1]. 이에 본 연구에서는 예능 콘텐츠에서도 STT 인식률이 저하되지 않도록 오디오 데이터에서 발화 음성 신호를 효과적으로 분리할 수 있는 딥러닝 기반의 음성 신호 필터를 제안한다. 또한, 제안 기술을 예능 콘텐츠의 오디오 데이터에 적용한 결과 실험을 진행함으로써, 제안 기술의 효용성을 검증하고자 한다.

Ⅱ. 본론

(1) 비음성 오디오(Instruments) 추출

본 연구에서는 U-Net 구조의 DNN 기반으로 학습된 네트워크(Vocal Remover)를 이용했다[2]. Vocal Remover 는 AR (All Recorded) 오디오 데이터로부터 음성신호(Vocal)가 제거된 MR(Music Recorded) 오디오 데이터를 추출하는 모델이다. 본 연구에서는 Vocal Remover 를 비음성 오디오 구간을 파악하기 위한 용도로 사용했다. 배경음악과 효과음이 포함된 예능 콘텐츠 오디오 데이터에 Vocal Remover 를 사용하여 배경음악과 효과음만 존재하는 비음성 오디오 데이터를 형성하고, 비음성 오디오가 존재하는 구간을 파악하였다.

(2) 데이터 생성과 모델 훈련

그림 1 은 제안 기술의 전체적인 데이터 전처리 과정을 나타낸다. Vocal Remover 를 통하여 추출한 비음성

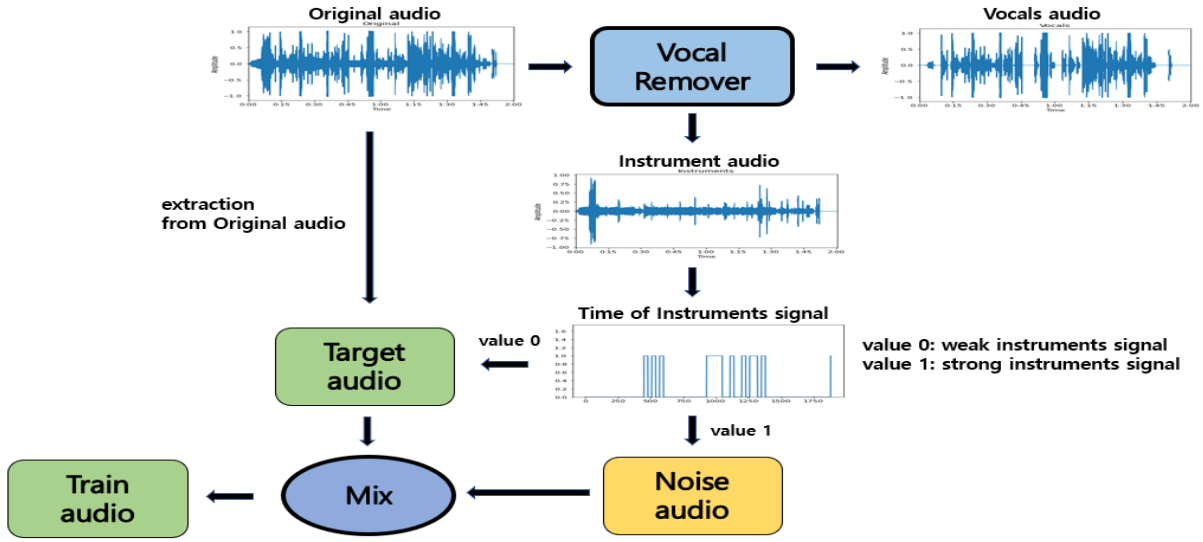

그림 1. 데이터 전처리 작업

오디오(Instruments)의 신호의 세기에서 적절한 임계치 설정을 통해 비음성 신호가 약한 구간을 결정한다. 그리고 원본 오디오에서 비음성 신호가 약한 구간들을 추출하여 새로운 오디오 데이터를 생성한다. 이 구간의 오디오 데이터는 음성 신호가 비음성 신호에 비해 강하게 존재하는 구간으로써 모델 훈련을 위한 정답(Target) 데이터로 설정한다. 반면, 비음성 신호가 강한 구간들을 모아서 Noise 오디오를 생성하고 Voice 오디오와 섞는 작업으로 훈련(Train) 데이터를 생성한다. 준비된 데이터를 이용하여 U-Net 구조의 DNN 기반 음성 신호 필터를 개발한다.

(3) STT 결과 비교

본 연구의 STT 결과를 비교하기 위해 Google Cloud Platform 의 Cloud Speech-to-Text API 를 사용했으며, 개발한 음성 신호 필터를 통해 얻은 Speech 오디오와 원본 오디오의 STT 결과를 비교해보았다. 표 1 은 STT 결과를 평균 신뢰도와 단어 개수를 정량적 척도로 삼아 비교해본 결과이다. 배경음악과 효과음이 포함된 예능 콘텐츠의 원본 오디오 데이터에 비해서 음성 신호 필터를 통해 얻은 Speech Data 로부터 더 많은 단어의 개수를 추출한 결과를 확인할 수 있다. 그러나 해당 결과에서 신뢰도가 다소 떨어지는 현상이 발생했다.

Ⅲ. 결론

본 연구에서 개발한 음성 신호 필터를 통해 추출된 오디오 데이터(Speech Data)가 원본 오디오 데이터에 비해 STT 로부터 평균적으로 더 많은 단어를 추출하는 것을 확인했고, 이는 STT 음성 인식 엔진이 해당 데이터의 언어적 요소를 효과적으로 인식하고 있다고 판단할 수 있다. 그러나, 본 연구에서 제안한 방법으로 얻어진 데이터는 STT 결과의 평균 신뢰도면에서는 다른 데이터에 비해 다소 떨어지는 모습을 확인할 수 있었으며 이에 대한 추가적인 연구가 필요하다. 또한, Google Cloud STT API 의 특성 상 한국어에 대한 텍스트 변환 정확도가 국내 STT API 에 비해 다소 떨어지는 현상을 고려할 때, 정확한 성능 검증을 위해 국내 STT API 에 대한 추가적인 실험이 필요할 것으로 보인다.

	Speech Separation		Original	
Content	Confidence	Words	Confidence	Words
1	0.799518	638	0.835153	517
2	0.798374	441	0.839062	424
3	0.802707	151	0.751183	138
4	0.795391	887	0.831401	720
5	0.779352	508	0.794423	476
6	0.754574	352	0.839071	345
7	0.789305	293	0.78287	278
8	0.823653	231	0.81037	229
9	0.842467	416	0.851178	380
10	0.881583	2641	0.86325	2597

표 1. STT 결과 비교

ACKNOWLEDGMENT

참 고 문 헌

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

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