비음성 신호의 분리를 위한 적응형 임계값 기반 음성 신호 필터

권준, 이성배, 박기범, 김규헌*
경희대학교
kjun1000@khu.ac.kr, rhee@khu.ac.kr, sesbgb@khu.ac.kr, *kyuheonkim@khu.ac.kr

Adaptive Threshold based Voice Signal Filter for separation of Instrument Audio

Joon Kwon, Seongbae Rhee, Kibeom Park, Kyuheon Kim*
Kyung Hee Univ.

요 약

GPU의 성능이 향상되면서 딥러닝 기술 또한 발전되어 다양한 딥러닝 기반 기술들이 이용되고 있다. 이는 음성 인식 기술에도 적용되어 기존 기술보다 더 나은 머신 태스크 성능을 보여주고 있다. 그 중, 딥러닝을 이용한 STT(Speech-to-Text) 기술은 입력으로 받은 음성 신호를 텍스트로 변환하는 기술로서 다양한 분야에서 이용되고 있다. 이러한 STT 기술은 뉴스와 같이 배경음이 적은 오디오 신호에 대해서는 비교적 높은 정확도를 보여주고 있지만, 예능 방송과 같이 사람의 음성과 배경음이 혼합된 오디오 신호에서는 낮은 정확도가 나타날 수 있다. 이에 본 논문에서는 배경음이 많이 섞여 있는 예능 방송 오디오 신호에 대해서도 STT 기술의 성능을 향상하기 위해 입력 오디오 신호를 Vocal Remover를 통해 음성 신호와 비음성 신호로 분리하고, 분리된 비음성 신호마다 특정 임계값을 설정하고 구간을 분리하여 음성 신호에서 제거되지 않은 잡음 신호를 처리하는 딥러닝 기반 음성 신호 필터를 제안하고자 한다. 또한, 임계값을 모든 데이터에 대해 동일하게 설정한 필터와 본 논문에서 제안한 음성 신호 필터를 적용한 신호를 상용 STT 기술에 적용한 결과를 비교함으로써, 제안 기술의 효능을 검증하고자 한다.

Ⅰ. 서론

딥러닝의 발전에 따라 STT(Speech-to-Text) 기술은 높은 머신 태스크 성능을 바탕으로 AI 챗봇, 인공지능 비서 등 다양한 서비스에서 활용되고 있다[1]. STT 기술은 입력으로 들어오는 음성 신호를 텍스트로 변환하는 기술로써, 음성 신호를 입력으로 특징(Feature)을 학습 및 분류하여 텍스트를 생성한다. 이때 모델의 학습에 활용된 음성 신호가 사람의 목소리만 존재하는 음성 신호였다면 뉴스와 같이 배경음이 적은 분야에 대해서는 높은 정확도를 보여줄 수 있지만[2], 배경음이 많이 섞여 있는 음성 신호가 입력으로 들어올 때는 정확도가 크게 저하될 수 있다.

이와 같은 제한 사항을 극복하기 위하여 딥러닝 기반 필터가 연구되었지만[3], 해당 방법은 필터의 비음성 신호를 분리하는 과정에서 모든 데이터에 대해 동일한 임계값을 사용하기에 각 비음성 신호의 특징이 무시될 수 있다는 제한 사항이 있다. 이에 본 논문에서는 비음성 신호마다 신호 특성에 맞게 임계값을 설정하는 딥러닝 기반 음성 신호 필터를 제안하고자 한다. 또한, 이전 연구의 임계값을 모두 동일하게 설정한 필터와 본 논문에서 제안한 비음성 신호마다 다른 임계값을 설정하는 필터의 STT 결과를 비교하여 제안 기술의 효능을 검증하고자 한다.

Ⅱ. 딥러닝 기반 음성 신호 필터

그림 1은 본 논문에서 제안하는 딥러닝 기반 음성 신호 필터의 구조도이다. 구조도에서는 가장 먼저 입력으로 들어오는 음성 신호를 Vocal Remover[4]를 이용하여 음성 신호(Vocal Audio)와 비음성 신호(Instrument Audio)로 나눈다. 하지만 Vocal Remover는 비음성 신호를 뽑아내는 것을 목적으로 만들어진 기술이기 때문에 추가적인 분리 과정이 필요하다. 이를 위해 본 논문에서는 여러 추가적인 모듈을 통해 배경음이 섞여 있는 오디오 신호에 대해서도 높은 STT 성능을 보여줄 수 있는 필터링 된 오디오 신호를 생성하고자 한다.

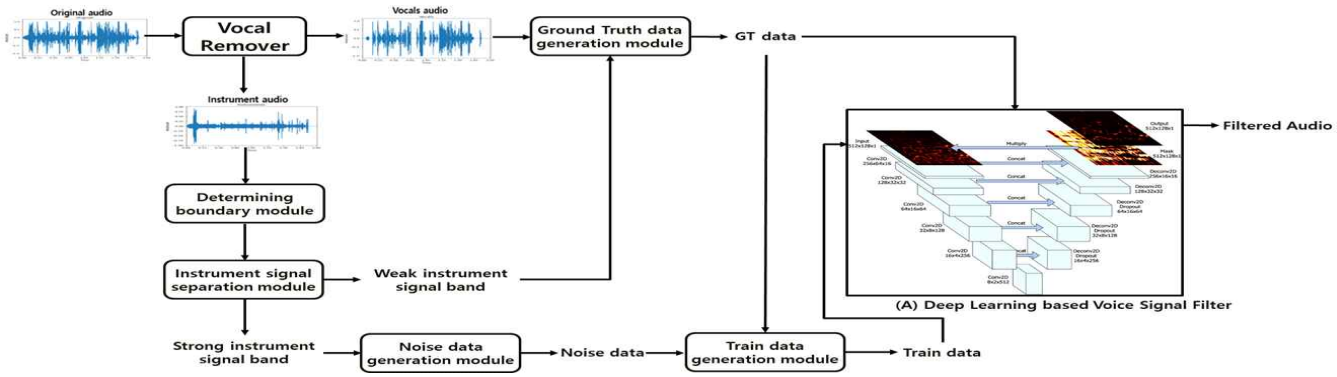
앞서 Vocal Remover를 이용하여 분리된 비음성 신호는 Determining boundary module을 통해 해당 비음성 신호의 특정 임계값이 정해지게 된다. 먼저, 비음성 신호를 N개로 샘플링하고 각 시간 축에서의 진폭 값의 평균을 구한다. 이후 계산된 평균값에서 특정 α 값을 빼 임계값을 정한다. 해당 내용에 대한 수식은 아래와 같다.

$$\theta_{th} = \frac{1}{N}\left(\sum_{i=1}^{N} \mid A_{inst}(i) \mid \right) - \alpha \quad \{\alpha \mid 0 < \alpha < 0.3\} \quad (1)$$

이후 Vocal Remover에서 분리된 비음성 신호는 특정 임계값을 기준으로 강한 비음성 신호 구간(Strong Instrument Signal band)과 약한 비음성 신호 구간(Weak Instrument Signal band)으로 분리된다. 분리된 약한 비음성 신호 구간은 해당 구간의 값들을 모두 1로 변환하고 그 외의 구간의 값들은 모두 0으로 변환하는 이진화 과정이 수행된다. 이후 이진화된 약한 비음성 신호 구간과 Vocal Remover에서 분리된 음성 신호와의 곱연산을 통해 모델 학습을 위한 GT(Ground Truth) data를 생성한다. 또한, 강한 비음성 신호 구간은 그대로 배경음에 해당하는 잡음 신호(Noise data)를 생성한다. 마지막으로 앞서 생성된 GT data와 Noise data가 합쳐져 Train data를 생성하게 된다. 생성된 Train data는 그림 1의 (A)에 표현된 딥러닝 기반 음성 신호 필터의 학습 데이터로 활용되며, 이때 활용하는 딥러닝 기반 음성 신호 필터는 그림 1의 (A)에서 보이듯이 U-net[5] 구조의 모델을 사용한다.

Ⅲ. 실험 결과

본 논문의 실험에서는 임계값을 모든 데이터에 대해 동일하게 설정한 필



<그림 1. 전처리 음성 신호 필터 구조도>

터와 데이터마다 임계값을 다르게 설정한 필터의 훈련을 진행하고, 해당 두 가지의 필터를 이용하여 원본 음성 신호와 두 가지의 필터를 각각 통과한 음성 신호의 STT 결과와 탐지된 단어 개수를 비교한다. 이때 STT 기술은 상용 Return Zero의 RTZR STT API[6]를 사용하였다.

	Original Audio		Filtered Audio (same threshold)		Filtered Audio (different threshold)		
Content	단어 개수	오류 비율	단어 개수	오류 비율	단어 개수	오류 비율	분류
1	315	0.193	301	0.183	304	0.170	스포츠
2	332	0.347	342	0.337	346	0.289	스포츠
3	229	0.256	228	0.235	226	0.214	드라마
4	139	0.268	148	0.318	124	0.272	드라마
5	313	0.355	284	0.423	262	0.344	예능
6	454	0.255	412	0.236	404	0.205	예능
7	427	0.288	415	0.283	417	0.248	예능
8	119	0.499	98	0.449	104	0.342	예능
9	190	0.220	182	0.185	184	0.132	예능
10	116	0.344	98	0.411	114	0.295	예능
11	246	0.293	218	0.306	216	0.274	예능
12	214	0.374	201	0.406	200	0.353	예능

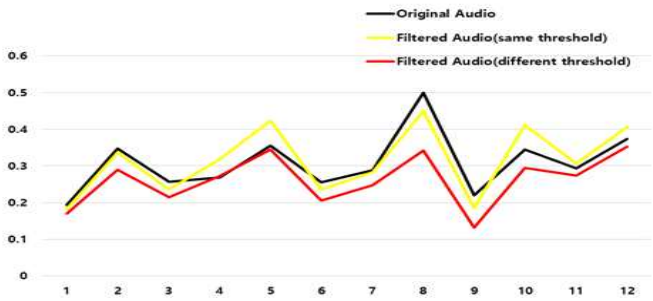
<표 1. STT 오류 비율, 단어 개수 비교>

표 1은 원본 오디오, 임계값이 동일한 필터를 통과한 오디오, 임계값이 데이터마다 다른 필터를 통과한 오디오의 STT 오류 비율과 탐지되는 단어 개수의 결과를 비교한 것이다. 세 가지 오디오의 STT 텍스트 결과에서 탐지되는 단어의 수는 원본 오디오, 임계값이 동일한 필터를 통과한 오디오, 본 논문에서 제안한 필터를 통과한 오디오 순으로 탐지되는 단어의 개수가 평균적으로 줄어드는 것을 표 1에서 확인할 수 있다. 이는 원본 오디오에서는 배경음에 해당하는 잡음 등을 모두 텍스트로 변환했지만, 제안 필터를 통과한 오디오에서는 해당 잡음 등이 효과적으로 제거되어 탐지되는 단어의 개수가 줄어든 것을 실험 결과에서 확인할 수 있었다.

그림 2는 원본 오디오, 임계값이 동일한 필터를 통과한 오디오, 임계값이 데이터마다 다른 필터를 통과한 오디오의 STT 오류 비율을 비교한 것이다. 그림 2에서 본 논문에서 제안한 필터를 통과한 오디오가 원본 오디오, 임계값이 동일한 필터를 통과한 오디오보다 평균적으로 더 좋은 성능을 보여주는 것을 확인할 수 있다. 또한, 표 1에서 빨간색으로 표현된 부분은 임계값이 동일한 필터를 통과한 오디오가 원본 오디오보다 STT 성능이 떨어지는 결과이다. 하지만 본 논문에서 제안한 필터를 통과한 오디오가 원본 오디오, 임계값이 동일한 필터를 통과한 오디오보다 평균적으로 더 좋은 성능을 보여주는 것을 통해 제안 필터의 효용성을 판단할 수 있다.

Ⅳ. 결론

본 논문에서는 예능 방송과 같이 배경음이 많이 섞여 있는 오디오에서 STT 성능이 저하되는 문제점을 극복하기 위한 딥러닝 기반 음성 신호 필터의 성능을 향상하기 위해, 비음성 신호의 임계값을 각 신호의 특성에 맞게 적용하여 필터를 학습시켰다.



<그림 2. STT 오류 비율 비교>

실험 결과를 통해 본 논문에서 제안한 필터를 통과한 오디오가 원본 오디오, 임계값이 모두 동일한 필터를 통과한 오디오보다 평균적으로 좋은 성능을 보여주는 것을 확인할 수 있었다.

다만, 이번 실험을 통해 오디오 신호를 나누고 합치는 과정에서 말이 끊기는 문제나, 부자연스러운 대화 등의 제한 사항들을 확인하였다. 따라서 향후 연구에서는 해당 문제들을 해결하여 양질의 데이터셋을 제작하는 방안에 대해서 진행할 필요성이 있다.

ACKNOWLEDGMENT

"This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-2021-0-02046) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)"

참 고 문 헌

[1] 박현신, 김성웅, 진민호, and 유창동, "최신 기계학습 기반 음성 인식기술 동향," 전자공학회지, Vol.41, No.3, pp.18-27, 2014.

[2] 황용해, 차은영, 홍순기, 김상진, 이학주, 서덕영, and 김규헌, "STT 정확도 향상을 위한 딥러닝 기반 MS 구간 추출 및 음성 분리," 한국통신학회 학술대회논문집, 개최지, pp.1559-1560, 2022.

[3] 권준, 이성배, 박기범, and 김규헌. "딥러닝을 이용한 음성 신호 필터". 한국방송미디어공학회 학술발표대회 논문집, 2023

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

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox "U-Net: Convolutional Networks for Biomedical Image Segmentation" (accessed May 18, 2015)

[6] returnzero, https://www.rtzr.ai/