직교 손실 함수를 통한 음성 분리 모델 성능 향상

우범준, 김정훈, 안성환, 전용현, 김남수
서울대학교
{bjwoo, jhkim, shahn, yhjeon}@hi.snu.ac.kr, nkim@snu.ac.kr

Enhancing Speech Separation model performance through Orthogonal loss function

Beom Jun Woo, Jeung Hun Kim, SungHwan Ahn, Yong Hyeon Jeon and Nam Soo Kim
Department of Electrical and Computer Engineering and INMC, Seoul National Univ.

요 약

본 논문은 기존 모델에 추가적인 모듈 및 입력 없이 성능을 올리는 방법을 제안한다. 특징 추출하는 feature가 분리가 더 잘 되도록 직교 함수를 이용한 loss function을 추가로 학습시켜 기존 모델 대비 음성 분리 모델이 학습을 더 잘할 수 있도록 제안한다. 2개의 딥러닝 기반의 음성 분리 모델에 적용하여 약 0.5dB의 성능 향상이 있음을 보인다.

Ⅰ. 서 론

딥러닝의 발전으로 인해 음성 분리 분야 또한 딥러닝 기반의 모델을 많이 사용하고 있다. 딥러닝 기반 음성 모델 구조에 있어, 혁신적으로 성능이 많이 증가하고 지금까지도 많이 사용되고 있는 구조의 근본이 되는 Dual Path RNN(DPRNN)[1]의 dual path 구조가 아직도 많이 사용되고 있다. 비슷한 구조 내의 변형을 통해 성능을 향상시키는 논문들이 많이 나오고 있지만, 모델 parameter가 증가한다는지, 추론 시간이 증가하는 등 여러 문제가 있다. 본 논문은 같은 모델 활용하에 추가적인 loss function을 이용하여 성능을 향상시키는 방법을 제안한다. 특징 feature이 구분이 쉽게 되도록 직교 함수의 특징을 활용하였다.

Ⅱ. 본론

1) 음성 분리 모델

본 논문의 음성 분리 모델로 DPRNN[1] 이라는 모델을 가져와서 사용한다. DPRNN은 encoder, separator, decoder 3가지 모듈로 이루어진 mask 기반의 음성 분리 모델이다. Encoder에서는 Shor time fourier transform 과 같은 형태로 학습 가능한 feature를 추출하게 된다. 추출한 feature는 separator 모듈에 들어가 짧은 구간의 특징과 긴 구간의 특징을 번갈아 가면서 학습을 진행하게 된다. Separator 모듈의 output은 사전에 지정된 화자의 수 만큼 input feature를 활용한 separation mask이다. Encoder에서 나온 feature와 separator module을 통해 나온 separation mask를 곱하면 2개의 feature가 생성된다. 2개의 feature를 decoder에 넣어주게 되면 화자별 음성이 복원된다.

2) 기존 Loss Function

기존 음성 분리 모델은 SI-SDR(Scale Invariant - Signal to Distortion Ratio)[2] 수식을 최대화하는 방향으로 학습이 진행된다. 수식은 아래와 같이 전개된다.

$$L_{SI-SDR} = 10\log_{10}\left(\frac{\left\| \frac{\hat{s}^T s}{\| s \|^2} s \right\|^2}{\left\| \frac{\hat{s}^T s}{\| s \|^2} s - \hat{s} \right\|^2}\right) \quad (1)$$

여기서 $\hat{s}$ 는 estimated speech이며 s는 reference speech이다. 기존 딥러닝 기반의 모델은 위 수식을 기반으로 각 화자에 대해 최대 값을 가지도록 학습을 진행한다.

3) 직교 특징 Loss Function

본 논문에서는 추출된 feature 간의 차이가 클수록 분리가 더 쉬울거라는 가정하에 그 차이를 극대화하기 위해 직교 특성을 가지도록 loss를 추가로 줬다. Encoder output을 E 라고 하였을 때 직교 특징 Loss 는 아래와 같이 구성된다.

$$L_o = \| EE^T - I \|^2 \quad (2)$$

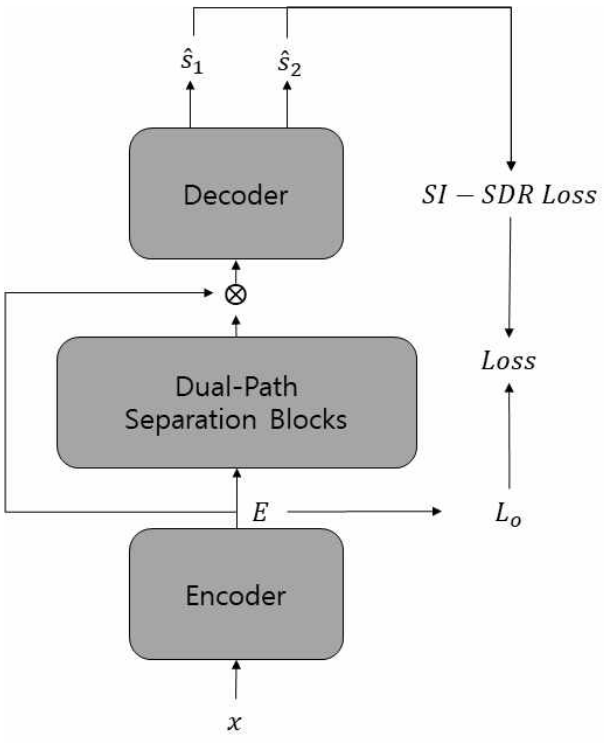

그림 1. Propose 하는 모델 학습 훈련 개요도

4) 실험 및 결과

본 논문에서는 Wall Street Journal 2mix(WSJ-2mix) data를 사용하여 학습을 진행한다. 본 데이터셋은 30시간의 훈련 데이터, 10시간의 검증 데이터 그리고 5시간의 테스트 데이터로 구성되어 있다.

총 Loss L은 그림 1의 개요도에서 나오는 것처럼 기존 SI-SDR loss와 직교 특징 loss를 더하여 최소화하는 방향으로 학습을 진행한다.

$$L = L_o - L_{SI-SDR} \quad (3)$$

본 모델은 기존 모델의 기본 구성과 다르게 dual path sepataion block을 6개에서 4개로 경량화하여 RTX 2080ti gpu에서 학습을 진행하였다.
실험 결과는 아래와 같다.

	SI-SDRi
Baseline	16.46dB
Proposed	**17.02dB**

표 1. DPRNN 실험 결과

이외에도 Conv-TasNet[3] 에도 실험을 똑같이 적용하여 진행해보았다.

	SI-SDRi
Baseline	14.86dB
Proposed	**15.31dB**

표 2. Conv-TasNet 실험 결과

Ⅲ. 결론

본 논문에서는 최근 나오고 있는 딥러닝 모델들의 추가적인 요소 없이 학습만으로 더 좋은 성능을 뽑고자 추가적인 직교 손실 함수에 관해 연구해보았다. 본 연구를 하는데 있어 encoder뿐만 아니라, decoder 및 mask를 입력 화자수 만큼 생성하는 부분 2 군데에도 추가적으로 본 논문과 같은 실험을 진행해 보았었는데 결과가 좋지는 못하였다. 결국 입력 feature를 처음부터 잘 분리해야 더 좋은 결과가 나오는 것을 확인할 수 있었다.

ACKNOWLEDGMENT

이 논문은 2024년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음.

참 고 문 헌

[1] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation,” in Proc. of ICASSP, 2020, pp. 46 - 50

[2] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr - half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pages 626 - 630. IEEE, 2019

[3] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing Ideal Time - Frequency Magnitude Masking for Speech Separation,” vol. 27, no. 8, pp. 1256 - 1266, Aug. 2019