화자의 특징 정보를 활용한 대상 음성 분리 알고리즘

우범준, 김석민, 김지환, 문성환, 윤지원, 이현승, 최병진, 한민현, 김남수
서울대학교
{bjwoo, smkim, jhkim21, shmun, jwyoon, hslee, bjchoi, mhhan}@hi.snu.ac.kr, nkim@snu.ac.kr

Targeted speech separation using speaker embedding

Beom Jun Woo, Seok Min Kim, Jihwan Kim, Sung Hwan Mun, Ji Won Yoon, Hyeon Seung Lee, Byoung Jin Choi, Min Hyun Han and Nam Soo Kim
Department of Electrical and Computer Engineering and INMC, Seoul National Univ.

요 약

본 논문은 지정된 화자의 음성 분리를 화자의 특징 정보를 활용하여 음성 분리 성능을 고도화하는 연구이다. 기존 음성 분리 네트워크와는 달리 원하는 화자를 지정하고 지정된 화자의 특징 정보를 벡터로 뽑아서 음성 분리 네트워크에 conditioning을 통해 성능을 고도화한다. 이 기법은 개인화 음성 향상, 잔향 제거와 같은 다른 downstream task 에도 쉽게 적용이 가능하며 음성 분리 네트워크에 적용을 통해 성능이 향상됨을 확인할 수 있다.

Ⅰ. 서 론

음성 분리란 2명 이상의 화자가 동시 발화하였을 때 음성이 섞여 개인의 발화가 무엇인지 인지하기 어려울 때 개별 음성으로 나누는 것이다. 최근 딥러닝의 발전으로 인해 다양한 음성 분리 알고리즘들이 등장하고 있지만, 대부분의 음성 분리 알고리즘들은 음성 분리 후 개별 음성을 다 내보낸다. 실제 환경에서의 인간은 2개 이상의 음성이 섞여 있는 것과 같이 주변 환경에 개의치 않고 자신과 관련된 정보만을 선택적으로 지각하여 잘 받아들인다. [1] 본 논문에서는 기존의 멀티 채널 음성 분리 알고리즘 [2]에 화자의 정보(speaker embedding)을 conditioning하여 특정 화자에 대한 음성 분리 성능을 고도화 할 수 있는 기법을 제안한다. 코드는 https://github.com/CARNIVAL-IITP/Speech_source_separation 에서 확인이 가능하다.

Ⅱ. 본론

1) 공간적 정보를 활용한 멀티 채널 음성 분리

Cone of Silence[2]란 멀티 채널 음성 분리 모델을 본 논문의 baseline으로 사용하였다.[그림1] 본 알고리즘은 $\{90^\circ, 45^\circ, 23^\circ, 12^\circ, 2^\circ\}$ 의 각도에 대해 화자의 목소리가 어디에 위치해 있는지 energy의 크기를 이용하여 특정 영역을 지정한다. 그리고 추정한 각도의 크기로 time differences of arrival (TDOA)를 측정하고, 혼합된 음성을 target wav와 align을 맞추도록 TDOA 값만큼 shift를 해준다. 전처리를 위와 같이 진행해주고 난 후 각각 2개의 convolution block으로 이루어진 u-net 구조로 학습을 진행하게 되면 원하는 위치에 속한 특정 화자 음성 분리를 진행하게 된다.

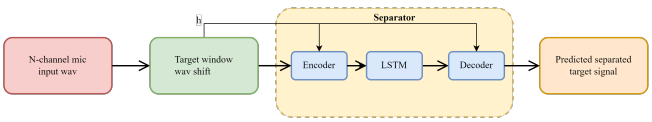

그림 1. baseline 훈련 모델

2) 공간 및 화자적 정보를 활용한 멀티 채널 음성 분리

혼합된 음성으로부터 특정 화자의 목소리를 더 잘 분리해내기 위해 모델에 화자의 정보를 conditioning 하여 음성 분리 성능을 고도화할 방법을 제안한다.

우선 화자의 정보는 특정 화자를 분리하려고 할 때 그 화자의 reference wav를 활용하여 화자의 정보를 추출한다. 화자의 정보를 추출하기 위해서는 pre-trained 된 ECAPA-TDNN[4]라는 모델을 활용한다.

conditioning 하는 방법은 기존 이미지에서 사용되었던 FiLM (Feature-wise Linear Modulation)[3]을 본 모델에 적용하여 학습을 진행한다. ECAPA-TDNN을 통해 추출된 화자의 정보 벡터를 λ_{spk} 라고 하고 기존 LSTM의 output을 h_{LSTM} 라고 하면 conditioning 수식은 아래와 같다.

$$h_{output} = h_{LSTM} \times lin_{mul}(\lambda_{spk}) + lin_{add}(\lambda_{spk}) \quad (1)$$

여기서 lin_{mul} 와 lin_{add}는 화자 정보 벡터를 fully connected layer을 뜻하며 h_{output}은 decoder의 input으로 들어가게 된다.

전체적인 모델의 흐름도는 아래의 그림 2와 같다. 아래의 그림은 그림 1의 baseline을 그대로 가져오면서 아래의 화자 정보 추출 네트워크와 화자 정보 벡터를 기존 멀티 채널 음성 분리 모듈에 컨디셔닝 하는 네트워크를 적용하였다.

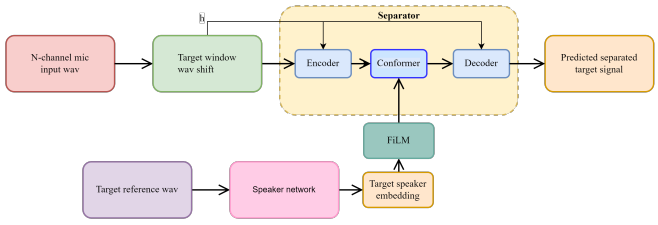

그림 2. Proposed 훈련 모델

3) 실험 및 결과

본 논문에서는 학습을 위해 사용한 음성 데이터는 Sitec DB로 각각 200명의 남자와 여자 화자로 구성되어있다. 본 모델의 훈련을 위해 남녀 각각 180명의 화자를 train dataset으로 구성하였고, 나머지 남녀 20명으로 validation 및 test dataset을 구성하였다. 멀티 채널 환경 구성을 위해서 실제 화상회의 환경과 비슷한 3가지 구성을 참고로 하여 RIR simulator를 활용하여 멀티 채널 데이터를 만들었다. 시뮬레이션 환경 구성은 표 1 과 같고 설치된 멀티 채널의 마이크의 구성은 표2와 같다.

환경 종류	중소형 회의실			중형 회의실			대형 회의실		
크기(m)	7.0	4.2	2.7	7.9	7.0	2.7	8.5	7.2	2.9
마이크 중심좌표(m)	3.5	2.1	0.7	3.95	3.5	0.7	4.25	3.6	0.7
RT60(s)	0.2			0.2			0.2		

표 1. 화의 환경

배열 종류	원형	타원형		선형
마이크(개)	반지름(cm)	장축(cm)	단축(cm)	길이(cm)
4	2.83	6.37	4.84	12
6	4	9.23	7.02	20
8	5.23	11.85	9.01	28

표 2. 마이크 구성

위와 같은 구성으로 train dataset을 원형, 타원형 그리고 선형 배열의 마이크들로 10000개를 구성하였고 test dataset을 방 구성마다 원형 400개 타원형 및 선형 300개로 구성하여 총 3천개로 구성하였다.

본 모델의 성능을 검증하는 데 있어 음성 분리에서 널리 쓰이는 지표 SI-SDRi(Scale Invariant - Signal to Distortion Ratio improvement)[5]를 사용하였다. 수식은 아래와 같다.

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

여기서 $\hat{s}$ 는 estimated speech이며 s는 reference speech이다.

멀티 채널 음성 분리 방법	4mic	6mic	8mic	Average
공간적 정보를 활용한 음성 분리 제거	6.72dB	7.45dB	8.24dB	7.47dB
공간 및 화자의 특징 정보를 활용한 음성 분리 제거	**8.49dB**	**9.42dB**	**10.14dB**	**9.29dB**

표 3. 실험 결과

표3의 결과를 통해 특정 화자를 분리하기 위해 화자 정보를 기존 네트워크에 conditioning하는 것이 성능을 향상하는 데에 큰 도움이 되는 것을 확인할 수 있었다. 표 4의 그림의 2.5초부터 3초까지의 spectrogram을 보면 다른 화자의 목소리를 제거해야 하는데 기존 방법은 다른 화자의 목소리가 남아있는 것을 확인 가능하였다.

Ⅲ. 결론

본 논문에서는 간단하면서 효율적인 화자 정보 conditioning을 통해 멀티 채널 음성 분리의 성능을 올릴 수 있는 것을 확인하였다. 하지만 그런데도 잔향 요소 및 잡음이 조금 남아있어 명료도를 개선할 방법이 더 필요하다. 또 모델의 실제로 활용하기엔 연산량이 많아서 이를 경량화할 필요가 있다.

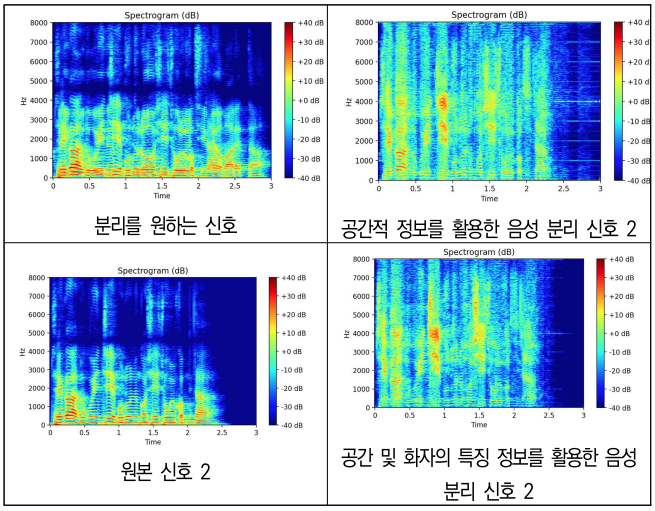

표 4. 각 방법에 따른 음성 분리 결과 spectrogram

ACKNOWLEDGMENT

이 논문은 2021년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음.

참 고 문 헌

[1] Wood N, Cowan N (January 1995). "The cocktail party phenomenon revisited: how frequent are attention shifts to one's name in an irrelevant auditory channel?". Journal of Experimental Psychology: Learning, Memory, and Cognition. 21 (1): 255 - 60.

[2] T. Jenrungrot, V. Jayaram, S. Seitz, and I.Kemelmacher Shlizerman, "The cone of silence: Speech separation by localization," in NeurIPS, 2020.

[3] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville. Film: Visual reasoning with a general conditioning layer. In AAAI, 2018

[4] B. Desplanques, J. Thienpondt, and K. Demuynk "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in Proc. Interspeech, 2020, pp. 3830-3834.

[5] 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.