# GEV 빔포밍을 위한 BiLSTM 기반 이진 마스크 추정

송일훈, 김홍국
광주과학기술원

ilhoon1204@gm.gist.ac.kr, hongkook@gist.ac.kr

# BiLSTM-Based Binary Mask Estimation for Generalized Eigenvalue Beamformer

Ilhoon Song, Hong Kook Kim
Gwangju Institute of Science and Technology

요 약

본 논문에서는 bidirectional long short-term memory (BiLSTM) 신경망을 학습하여 generalized eigenvalue (GEV) 빔포밍을 위한 이진 마스크를 추정한다. 그리고 이를 GEV 빔포머의 가중치를 구하는 과정과 후처리 과정에 각각 적용하여 기존 BiLSTM 기반 GEV 빔포머와 비교해 본 결과, perceptual evaluation of speech quality 와 speech-to-distortion ratio 수치에서 각각 12.63%와 15.06%가 향상됨을 확인하였다.

## Ⅰ. 서 론

빔포밍(Beamforming)은 잡음 환경에서 입력된 다채널 신호 각각에 특정 가중치를 주어 가중합을 구하는 방식으로 목표 음원 방향의 음질을 향상시키는 기법이다. 빔포밍 기법 중에서는 minimum variance distortionless response (MVDR)[1]과 generalized eigenvalue (GEV) 빔포밍[2]이 주로 연구되고 있다. MVDR 빔포머는 목표 음원 방향의 이득은 1 로 유지하면서 출력 잡음의 크기를 최소화한다. GEV 빔포머는 여러 음원 방향에 대한 신호 대 잡음 비를 구하여 이중 가장 큰 값을 찾으며 이를 토대로 목표 음원을 향상시킨다. 최근에는 딥러닝 기술이 발전함에 따라 빔포머 가중치를 구하는 과정에서 인공 신경망 구조를 사용하고 있으며 현재까지 bidirectional long short-term memory (BiLSTM) 신경망 구조를 이용한 GEV 빔포머가 높은 성능을 보이고 있다[3].

본 논문에서는 BiLSTM 신경망을 학습하여 이진 마스크(ideal binary mask)를 추정하고, 이를 통해 GEV 빔포머 가중치를 구한다. 또한, 가중치를 구한 이후에 후처리 과정으로 음성에 대한 이진 마스크를 적용하여 perceptual evaluation of speech quality (PESQ), speech-to-distortion ratio (SDR) 수치로 성능을 평가한다.

## Ⅱ. 본론

### 2.1 GEV 빔포머 모델

음성과 잡음이 혼합되어 있는 다채널 입력 신호는 다음 식과 같이 정의할 수 있다.

$$\mathbf{Y}_{f,t} = \mathbf{X}_{f,t} + \mathbf{N}_{f,t} \tag{1}$$

여기서, $\mathbf{Y}_{f,t}$는 short-time Fourier transform (STFT)를 적용한 다채널 입력 신호, $\mathbf{X}_{f,t}$는 음성 신호, $\mathbf{N}_{f,t}$는 잡음 신호이다. $f$는 주파수이며 $t$는 음성 프레임을 의미한다. GEV 빔포머 가중치 $\mathbf{w}_f$ 는 [2]에 의해 다음 식으로 표현된다.

$$\mathbf{w}_f = \underset{\mathbf{w}_f}{\operatorname{argmax}} \frac{\mathbf{w}_f^H \mathbf{\Phi}_f^{(X)} \mathbf{w}_f}{\mathbf{w}_f^H \mathbf{\Phi}_f^{(N)} \mathbf{w}_f} \tag{2}$$

여기서, $\mathbf{\Phi}_f^{(X)}$와 $\mathbf{\Phi}_f^{(N)}$은 각각 음성과 잡음에 대한 공분산 행렬이며 고유값은 아래 식으로 구할 수 있다.

$$\left\{\mathbf{\Phi}_f^{-(N)} \mathbf{\Phi}_f^{(X)}\right\} \mathbf{w}_f = \boldsymbol{\lambda} \mathbf{w}_f \tag{3}$$

고유값 분해식에 의해 여러 고유값 중 가장 큰 고유값에 해당하는 빔포밍 가중치가 결정되며 이는 여러 음성 발화 방향에 대해 신호대잡음비를 구하여 최종 목표 음원 방향을 추정하는 것이다. 또한, 음성과 잡음에 대한 공분산 행렬은 다채널 입력 신호와 음성 성분의 마스크 $M_{f,t}^{(X)}$ 및 잡음 성분의 마스크 $M_{f,t}^{(N)}$로부터 얻어질 수 있다.

$$\mathbf{\Phi}_f^{(X)} = \sum_{t=1}^{T} M_{f,t}^{(X)} \mathbf{Y}_{f,t} \mathbf{Y}_{f,t}^H , \tag{4}$$

$$\mathbf{\Phi}_f^{(N)} = \sum_{t=1}^{T} M_{f,t}^{(N)} \mathbf{Y}_{f,t} \mathbf{Y}_{f,t}^H . \tag{5}$$

본 논문에서는 음성 성분의 마스크를 1, 잡음 성분의 마스크를 0 으로 판단하는 이진 마스크를 사용하며 이러한 마스크는 BiLSTM 신경망 기반으로 훈련된다. 이진 마스크 훈련 시, 실제 정답 값과 가장 가까운 값이 예측되도록 비용 함수로써 binary cross entropy (BCE)를 사용한다. 훈련 목표 값으로 정답 음성에 대한 이진 마스크와 신경망의 출력에 대한 이진 마스크를 비교하여 손실을 최소화 하도록 한다.

그림 1 은 BiLSTM 을 이용하여 이진 마스크가 훈련되는 전체적인 과정을 보여 준다. 그림 1(a)는 특징 추출 과정으로써 다채널 입력 신호를 STFT 적용 하여 513 주파수 빈으로 만들고 이를 BiLSTM 입력으로 준다. 그림 1(b)는 신경망 훈련 과정으로써 BiLSTM 층 2 개와 전결합 층(fully-connected layer) 2 개, 이후 sigmoid

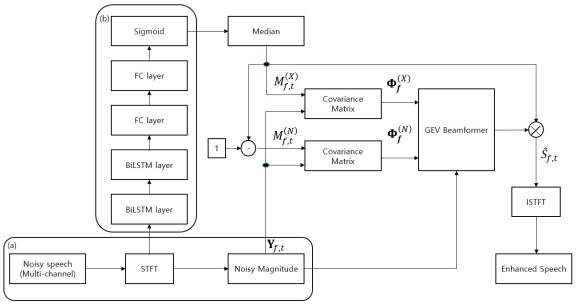


**그림 1.** BiLSTM 을 이용한 이진 마스크 추정 모델; (a) 특징 추출 과정, (b) 신경망 훈련 과정

함수를 거쳐 음성에 대한 다채널 이진 마스크를 생성한다. 다채널 이진 마스크는 median 연산을 통해 단일 채널 이진 마스크로 생성되며 잡음에 대한 이진 마스크는 $(1 - M_{f,t}^{X})$ 연산하여 생성한다. GEV 가중치를 구한 이후에는 다채널 입력신호에 대한 가중합으로 나타낼 수 있다. 여기서 후처리 과정으로 음성에 대한 이진 마스크를 적용하여 잡음을 최소화할 수 있으며 이렇게 향상된 음성에 대한 추정치 $\hat{S}_{f,t}$ 는 다음 식과 같다.

$$\hat{S}_{f,t} = \mathbf{w}_f^H \mathbf{Y}_{f,t} \cdot M_{f,t}^{(X)} \quad (6)$$

이후, inverse STFT 을 통해 신호를 주파수 축에서 시간 축으로 변환하며 최종적으로 향상된 단일 채널의 음성 신호를 얻을 수 있다.

### 2.2 실험 및 성능 평가

실험을 위해 CHiME-3 7,138 개 발화 문장으로 구성된 Bus, Cafe, Pedestrian area, Street 잡음 환경에서의 훈련 데이터로 학습을 진행하였다[4]. 테스트 과정에서는 1,320 개 발화 문장으로 구성된 테스트 데이터로 평가하였다. 성능 평가 결과, 기존 BiLSTM 기반 GEV 빔포머(BiLSTM-GEV)에 후처리 과정으로 음성에 대한 이진 마스크를 적용한 모델(BiLSTM-GEV+IBM)이 잡음을 최소화하여 각각 모든 잡음 환경에 대해 PESQ, SDR 수치를 평균적으로 각각 12.63%, 15.06% 향상시킬 수 있었다.

**표 1.** GEV 빔포머 모델 별 PESQ, SDR score 비교

| Noise Type | Model | PESQ | SDR (dB) |
|---|---|---|---|
| Bus | Noisy data (6-ch) | 1.71 | 0.28 |
| | BiLSTM-GEV | 2.88 | 5.70 |
| | BiLSTM-GEV+IBM | 3.14 | 8.14 |
| Cafe | Noisy data (6-ch) | 1.51 | 1.15 |
| | BiLSTM-GEV | 2.60 | 7.01 |
| | BiLSTM-GEV+IBM | 3.01 | 7.20 |
| Pedestrian | Noisy data (6-ch) | 1.50 | 1.26 |
| | BiLSTM-GEV | 2.68 | 7.03 |
| | BiLSTM-GEV+IBM | 3.04 | 7.14 |
| Street | Noisy data (6-ch) | 1.51 | 0.63 |
| | BiLSTM-GEV | 2.69 | 6.92 |
| | BiLSTM-GEV+IBM | 3.02 | 7.83 |

## Ⅲ. 결론

본 논문에서는 GEV 빔포밍을 위한 BiLSTM 기반 이진 마스크 추정 기법을 소개하고, GEV 빔포머의 가중치를 구하는 과정과 후처리 과정에서 각각 이진 마스킹을 적용하였다. 그리고 이를 CHiME-3 데이터셋을 사용하여 성능 평가한 결과, PESQ, SDR 수치에서 보다 높은 성능을 보이는 것을 확인할 수 있었다.

## ACKNOWLEDGMENT

본 연구는 2020 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구이며(No. 2019-0-01767, 드론을 활용한 재난 대응을 위한 기계학습 기반 음향지능 기술 개발), 그리고 2020 년도 광주과학기술원 GRI (GIST 연구원)의 지원을 받아 수행된 연구임.

## 참 고 문 헌

[1] C.-Y. Chen and P. P. Vaidyanathan, “Quadratically constrained beamforming robust against direction-of-arrival mismatch,” *IEEE Trans. Signal Process.*, vol. 55, no. 8, pp. 4139-4150, Aug. 2007.

[2] E. Warsitz and R. Haeb-Umbach, “Blind acoustic beamforming based on generalized eigenvalue decomposition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1529-1539, July 2007.

[3] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, “BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge,” *IEEE 2015 Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 444-451, 2015.

[4] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ speech separation and recognition challenge: dataset, task and baselines,” in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 504-511, 2015.