논문 21-46-06-18 The Journal of Korean Institute of Communications and Information Sciences '21-06 Vol.46 No.06
https://doi.org/10.7840/kics.2021.46.6.1078

일반화된 고유값 빔형성을 위한 양방향 장단기 메모리 기반 마스크 후처리 기법

송 일 훈◆, 김 홍 국°

BiLSTM-Based Mask Post-Processing Method for a Generalized Eigenvalue Beamformer

Ilhoon Song◆, Hong Kook Kim°

요 약

일반화된 고유값(generalized eigenvalue, GEV) 빔형성은 다채널 마이크로폰 어레이 구조에 독립적으로 도래각 추정에 의존하지 않으면서 잡음 환경에서의 음성을 추정할 수 있다. GEV 빔형성을 위해서는 타겟 음성 및 잡음의 전력 스펙트럼 밀도 행렬을 구해야 한다. 본 논문에서는 GEV 빔형성을 위한 양방향 장단기 메모리(bidirectional long short-term memory, BiLSTM) 신경망 기반 이진 마스크 후처리 기법을 제안한다. 제안된 BiLSTM은 다채널 잡음 음성의 스펙트로그램을 입력으로 하고 이진 마스크를 타겟으로 하여 학습되며, BiLSTM으로 추정된 이진 마스크를 적용하여 음성 및 잡음의 전력 스펙트럼 밀도 행렬을 구하고, 이를 이용하여 고유값 분해를 통해 GEV 가중치를 추정한다. 또한, BiLSTM 기반으로 추정된 이진 마스크를 GEV 빔형성으로 처리된 음성에 추가적으로 적용되어 clean 음성 추정 성능 개선에 활용된다. 제안된 방법의 성능을 평가하기 위하여 CHiME-3 데이터셋에 적용하여 실험한 결과, 기존의 BiLSTM 기반 이진 마스크 추정을 GEV 빔형성에 적용한 경우와 비교하여, 제안된 방법이 perceptual evaluation of speech quality (PESQ)에서 0.34 mean opinion score (MOS)와 signal-to-distortion ratio (SDR)에서 0.91 dB를 개선하였다.

Key Words : Generalized Eigenvalue (GEV) Beamforming, Speech Enhancement, Bidirectional Long Short-Term Memory(BiLSTM), Ideal Binary Mask Estimation, Post-Processing

ABSTRACT

Generalized eigenvalue (GEV) beamforming can estimate a target speech signal from a multi-channel microphone array in noisy environments, where it does not rely on the array structure as well as direction-of-arrival (DOA) of the target speech. The GEV beamformer is realized by using power spectrum density (psd) matrices for the target speech and noise. This paper proposes a binary mask post-processing method based on a bidirectional long short-term memory (BiLSTM) neural network for GEV beamformer. The BiLSTM is trained by using spectrograms of multi-channel input speech and ideal binary mask as input and

※ 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2021-0-00014, 재난상황 대응을 위한 엣지컴퓨팅 기반 시청각 인지지능 솔루션 개발)과 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2019-0-01767, 드론을 활용한 재난 대응을 위한 기계학습 기반 음향지능 기술 개발)

◆ First Author : Gwangju Institute of Science and Technology, School of Electrical Engineering and Computer Science, ilhoon1204@gm.gist.ac.kr, 학생회원

° Corresponding Author : Gwangju Institute of Science and Technology, School of Electrical Engineering and Computer Science/ AI Graduate School, hongkook@gist.ac.kr, 종신회원

논문번호 : 202103-049-A-RE, Received March 4, 2021; Revised April 6, 2021; Accepted April 7, 2021

outputs, respectively. Then, the estimated binary mask is applied to multi-channel noisy speech signals to obtain the speech and noise psd matrices. To further improve the quality of enhanced speech, the estimated binary mask is also applied to a post-processing stage of the GEV beamformer. The performance of the GEV beamformer is evaluated on a task of CHiME-3 by measuring the perceptual evaluation of speech quality (PESQ) and signal-to-distortion ratio (SDR). Experiments show that the GEV beamformer employing the proposed binary mask post-processing method improves PESQ and SDR by 0.34 mean opinion score (MOS) and 0.91 dB, respectively, compared to the conventional BiLSTM-based mask estimation method.

Ⅰ. 서 론

최근, 많은 양의 음성 데이터의 확보와 함께 딥러닝(deep learning) 기반의 음성처리 기법들에 관한 연구가 활발히 진행되고 있다. 특히, 음성처리의 응용으로 음성인식 분야 또한 꾸준한 발전을 이루며 사용자들에게 편리함을 제공하고 있다. 하지만, 일상생활 환경에서는 주변에 다양한 잡음이 존재하고, 마이크로폰을 통한 음성을 수집할 때 원하지 않는 잡음이 함께 녹음되어 음성인식 성능이 저하되는 문제점을 지니고 있다[1].

이러한 문제를 해결하기 위해 음성향상 기술이 등장하였다. 음성향상은 음성에 포함된 잡음에 의한 영향을 제거하는 기술로써 음질이나 음성의 명료도를 높이는 것을 의미한다. 전통적인 단일 마이크로폰 기반의 음성향상 기법과 더불어, 최근에는 하드웨어의 신호처리 연산 능력이 향상되면서 다양한 스마트 기기들이 여러 개의 마이크로폰을 사용하는 복잡한 신호처리가 가능해졌다. 따라서 다중 마이크로폰 기반의 음성향상 기술이 단일 마이크로폰을 사용하는 음성향상 기술보다 더욱 효과적으로 잡음을 제거할 수 있다[2].

2개 이상의 다중 마이크로폰을 이용한 다채널 음성향상 기술의 경우 빔형성(beamforming) 기술이 현재 널리 사용되고 있다. 빔형성 기술은 음성인식뿐만 아니라 스마트 자동차, 스마트 홈 케어 시스템 등과 같은 잡음 환경에서 동작하는 음성 기능을 갖는 시스템에 다양하게 활용되고 있으며, 최근에는 보청기, 헤드폰 및 원격 화상 회의 시스템 등에서도 사용되고 있다[3]. 빔형성은 잡음 환경에서 입력된 다채널 신호 각각에 특정 가중치를 주어 가중합을 구하는 방식인데, 이는 음성이 존재하는 방향의 소리를 증폭하고 다른 방향의 소리는 감쇄시키는 필터로 목표 음원 방향의 신호만을 추출하는 방법이다. 다양한 빔형성 기법 중에서는 최소 분산 비왜곡 응답(minimum variance distortionless response, MVDR)[4]과 일반화된 고유값(generalized eigenvalue, GEV)[5] 빔형성이 주로 연구되고 있다. MVDR 빔형성은 목표 음원 방향의 이득은 1로 유지하면서 출력 잡음의 크기를 최소화하는 기법이다. 또한, MVDR 빔형성 가중치는 조향응답파워 위상변환(steered-response power phase transform, SRP-PHAT)[6] 알고리즘을 통해 얻은 도래각(direction-of-arrival, DOA) 추정으로 계산할 수 있다.

이에 반해 GEV 빔형성 기법은 각 주파수 빈(frequency bin)에서 빔형성 출력의 신호 대 잡음비(signal-to-noise ratio, SNR)를 최대화하는 값을 찾으며 이를 토대로 목표 음원을 향상한다[5]. 이러한 빔형성 기법은 마이크로폰 어레이 구조에 독립적이며 DoA에 대한 명확한 추정을 필요로 하지 않는다[5].

MVDR이나 GEV 빔형성은 전력 스펙트럼 밀도(power spectral density)를 계산하여 가중치를 구하게 된다. 이때 음원에서 잡음에 대한 영향이 큰 경우, MVDR 빔형성은 DoA에서 오차가 발생하며 이로 인해 전력 스펙트럼 밀도를 계산하기 어렵다는 단점이 존재한다. 반면, GEV 빔형성은 DoA에 대한 영향을 받지 않기 때문에 CHIME-3[7] 음원 데이터셋과 같이 SNR이 낮은 환경 조건에서도 전력 스펙트럼 밀도를 계산할 수 있다.

GEV 빔형성 가중치는 구해진 음성 및 잡음의 전력 스펙트럼 밀도 행렬을 토대로 일반화된 고유값 문제를 해결함으로써 결정된다. 또한, 이러한 전력 스펙트럼 밀도 행렬은 다채널 음성 신호에 이진 마스크를 적용하여 얻어질 수 있다[8]. 다채널 음성 신호에 음성 이진 마스크를 적용하면 음성에 대한 전력 스펙트럼 밀도 행렬이 구해지며 반대로 잡음 이진 마스크를 적용하면 잡음에 대한 전력 스펙트럼 밀도 행렬을 구할 수 있다. 이는 이진 마스크를 통해 전력 스펙트럼 밀도를 계산할 수 있다는 것을 의미하며 결국, 음성 및 잡음 성분에 대한 이진 마스크가 얼마나 잘 추정되었는가에 따라 GEV 빔형성의 성능이 좌우될 수 있다.

따라서, 본 논문에서는 음성 및 잡음 성분에 대한 이진 마스크를 효과적으로 추정하여 GEV 빔형성의 성능을 높일 수 있도록 양방향 장단기 메모리

(bidirectional long short-term memory, BiLSTM)[9] 신경망 기반 이진 마스크 후처리 기법을 제안한다. 먼저 제안된 기법은 음성 및 잡음 성분에 대한 이진 마스크를 생성할 수 있도록 BiLSTM 신경망을 이용하며 다채널 음성 신호를 해당 신경망의 입력으로 한다. 다음으로 향상된 음원을 얻기 위해 BiLSTM 신경망 출력으로부터 생성되는 음성 및 잡음에 대한 이진 마스크를 GEV 빔형성 가중치를 계산하는 과정과 GEV 빔형성 출력단계에서 각각 적용한다.

본 논문의 구성은 다음과 같다. Ⅱ절에서는 기존의 GEV 빔형성 기반 음성향상 기법을 소개한다. Ⅲ절에서는 GEV 빔형성을 위한 마스크 후처리 기법을 제안하고 이를 이용한 음성향상 기법을 기술한다. 그리고 Ⅳ절에서는 제안 기법에 대한 성능을 평가한다. 마지막으로, Ⅴ절에서는 결론을 맺는다.

Ⅱ. 기존의 GEV 기반 빔형성 기법

2.1 일반적인 빔형성 모델

그림 1은 다채널 음성 신호에서 빔형성 가중치를 가중 합하여 향상된 음성을 얻는 과정을 보여주고 있다. 그림에서 보는 바와 같이, 2채널 이상의 다채널 입력 신호 $\mathbf{y}(t)$는 다음과 같이 정의할 수 있다.

$$\mathbf{y}(t) = \left[y_1(t), y_2(t), \cdots, y_M(t)\right] \tag{1}$$

여기서, $y_M(t)$는 시간 t에서 M번째 채널로 녹음되는 신호를 의미한다. 일반적인 빔형성 가중치 $\mathbf{w}$는 다음과 같이 표현될 수 있다.

$$\mathbf{w} = \left[w_1, w_2, \cdots, w_M\right] \tag{2}$$

여기서, w_M은 M번째 채널로 녹음되는 빔형성 가중치를 의미한다. 가중치를 어떻게 결정할 것인가에 따라 빔형성의 종류가 나뉜다. 이후, 가중치와 다채널 신호

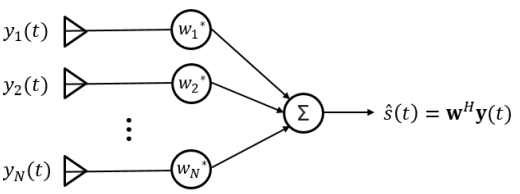

그림 1. 다채널 음성에서 향상된 음성을 얻는 과정
Fig. 1. Process of obtaining enhanced speech from multi-channel speech

를 가중 합하여 향상된 음성 추정치 $\hat{s}(t)$는 다음과 같다.

$$\hat{s}(t) = \mathbf{w}^H\mathbf{y}(t) \tag{3}$$

여기서, H는 Hermitian operator를 의미한다.

2.2 GEV 빔형성 모델

다음으로 GEV 빔형성 모델을 소개한다. 먼저 음성과 잡음이 혼합된 다채널 입력 신호는 다음 식과 같이 정의할 수 있다.

$$\mathbf{Y}_{t,f} = \mathbf{X}_{t,f} + \mathbf{N}_{t,f} \tag{4}$$

여기서, $\mathbf{Y}_{t,f}$는 단시간 푸리에 변환(short-time Fourier transform, STFT)을 적용하여 시간-주파수 도메인으로 변환한 신호이다. 그리고 $\mathbf{X}_{t,f}$는 깨끗한(clean) 음성 신호, $\mathbf{N}_{t,f}$는 잡음 신호이다. 또한 f는 주파수이며 t는 음성 프레임을 의미한다. GEV 빔형성 가중치 $\mathbf{w}_f$는 식 (5)와 같이 표현될 수 있다.

$$\mathbf{w}_f = \underset{\mathbf{w}_f}{\arg\max} \frac{\mathbf{w}_f^H \boldsymbol{\Phi}_f^{(X)} \mathbf{w}_f}{\mathbf{w}_f^H \boldsymbol{\Phi}_f^{(N)} \mathbf{w}_f} \tag{5}$$

여기서, $\boldsymbol{\Phi}_f^{(X)}$와 $\boldsymbol{\Phi}_f^{(N)}$은 각각 음성과 잡음에 대한 전력 스펙트럼 밀도 행렬이다. GEV 빔형성은 각 주파수 빈에서 SNR을 최대화하는 것이다. 이는 다음과 같은 일반화된 고유값 문제로 정의할 수 있다.

$$\left\{\boldsymbol{\Phi}_f^{-(N)} \boldsymbol{\Phi}_f^{(X)}\right\} \mathbf{W} = \lambda \mathbf{W} \tag{6}$$

여기서, $\mathbf{W}$는 고유벡터(eigenvector), λ는 고유값(eigenvalue)에 해당한다. 일반화된 고유값 분해 식에 의해 여러 고유값 중 가장 큰 고유값에 해당하는 고유벡터가 결정되고 이 고유벡터가 빔형성 가중치로 정의된다.

또한, 음성과 잡음에 대한 전력 스펙트럼 밀도행렬은 다채널 입력 신호인 $\mathbf{Y}_{t,f}$에 음성 성분의 마스크, $M_{t,f}^{(X)}$ 및 잡음 성분에 대한 마스크, $M_{t,f}^{(N)}$를 적용하여 얻어질 수 있다.

$$\boldsymbol{\Phi}_f^{(X)} = \sum_{t=1}^{T} M_{t,f}^{(X)} \mathbf{Y}_{t,f} \mathbf{Y}_{t,f}^H \tag{7}$$

$$\boldsymbol{\Phi}_f^{(N)} = \sum_{t=1}^{T} M_{t,f}^{(N)} \mathbf{Y}_{t,f} \mathbf{Y}_{t,f}^H \tag{8}$$

여기서, T는 음성 프레임의 총 개수이다.

한편, 잡음 제거와 음성 왜곡은 반비례 관계를 지니고 있으므로 GEV 빔형성 과정에서 SNR을 최대화하게 되면 음성 왜곡을 발생시킬 수 있다[10]. 이러한 음성 왜곡을 제거하기 위해서 단일 채널 후처리 필터(post-filter)인 블라인드 분석 정규화(blind analytic normalization, BAN)[5]를 사용한다.

$$\omega_f^{(BAN)} = \frac{\sqrt{\mathbf{w}_f^H \boldsymbol{\Phi}_f^N \boldsymbol{\Phi}_f^N \mathbf{w}_f / M}}{\mathbf{w}_f^H \boldsymbol{\Phi}_f^N \mathbf{w}_f} \tag{9}$$

다음으로, 다채널 신호와 빔형성 가중치에 BAN 필터까지 적용하여 향상된 음성 추정치 $\hat{S}_{f,t}$는 다음과 같이 정의할 수 있다.

$$\hat{S}_{t,f} = \omega_f^{(BAN)} \mathbf{w}_f^H \mathbf{Y}_{t,f} \tag{10}$$

Ⅲ. GEV 빔형성을 위한 마스크 후처리 기법

3.1 이진 마스크 생성 과정

GEV 빔형성을 위해서는 Ⅱ절에서 기술한 바와 같이, 식 (7)과 (8)의 전력 스펙트럼 밀도 행렬을 계산하여야 한다. 본 논문에서는 음성과 잡음 성분을 분리하기 위해 이진 마스크(ideal binary mask, IBM)[11]를 생성하고 이를 통해 음성과 잡음의 전력 스펙트럼 밀도 행렬을 구한다. 이진 마스크는 시간-주파수 스펙트럼의 음성 및 잡음 부분만을 출력하도록 마스킹한다. 이러한 마스크 적용으로 인해 GEV 빔형성 가중치는 잡음의 공간적 상관관계에 관한 가정을 필요로 하지 않는다. 그 이유는 목표 음원이 전력 스펙트럼 밀도 행렬 $\boldsymbol{\Phi}_f^{(X)}$에서, 잡음에 대한 음원이 $\boldsymbol{\Phi}_f^{(N)}$에서 집중되어 있다는 가정 때문이다[5].

이진 마스크는 시간-주파수 도메인에서 다음 식과 같이 구할 수 있다.

$$\mathrm{IBM} = \begin{cases} 1, \text{ if } SNR_{t,f} > \theta \\ 0, \text{ otherwise} \end{cases} \tag{11}$$

여기서, $SNR_{t,f}$가 임계값 θ를 초과할 경우 IBM은 '1'의 값을 가지고 그렇지 않으면 '0'을 갖는다[11]. 본 논문에서는 기존 연구[12,13]을 기준으로 임계값 θ를 0 dB로 설정하였다.

3.2 BiLSTM 신경망 기반 이진 마스크 추정 과정

음성 및 잡음에 대한 이진 마스크는 BiLSTM 신경망으로 학습한다. 이진 마스크 훈련 시, 실제 정답 값과 가깝게 예측되도록 이진 교차 엔트로피(binary cross entropy) 손실 함수를 사용한다. Clean 다채널 음성의 마스크를 훈련 목표 마스크로 설정한 뒤 음성과 잡음이 혼합된 다채널 신호에 대한 마스크를 모델의 예측값으로 정하고 훈련을 진행한다. 이후, 목표 마스크와 예측 마스크를 비교하여 손실을 최소화하는 방향으로 학습이 진행된다.

그림 2는 GEV 빔형성을 위한 제안된 이진 마스크 추정 기법의 전체적인 구성도를 보여준다. 특히, 그림 2의 점선 블록은 특징 추출 및 이진 마스크에 대한 훈련과정이며, 그림 3은 이 부분을 자세히 보여준다. 그림 3(a)는 모델 훈련을 위한 특징 추출 단계를 보여주고 있다. 이 단계에서는 먼저 다채널 음성 신호에 대해 STFT를 적용한다. 이 과정을 통해 음성 신호를 513 주파수 빈으로 만들고 이를 BiLSTM 신경망의 타임 스텝(time step)마다 입력한다. 그리고 해당 신경망 모델은 그림 3(b)에서와 같이 BiLSTM층 2개와 전결합층(fully-connected layer) 2개, sigmoid 활성화 함수로 이루어져 있다. 신경망 학습 시에는 먼저 가중치 초기화를 위해 Xavier uniform initialization[14]을 사용한다. 은닉층에는 비선형 변환으로 정류 선형 유닛(rectified linear unit, ReLU)을 이용하고 BiLSTM 각 층에 배치 정규화[15]를 사용한다. 그리고 모델의 과적합이 발생하는 것을 막기 위해 dropout(=0.5)을 각 층 마다 적용한다[16]. 또한, 안정적인 훈련을 위해서 Adam (adaptive moment estimation) 최적화를 채택했고 학습률은 0.001로 설정하였다.

다음으로 각 신경망 층마다 연결되는 출력 노드 수를 기술한다. 첫 번째 BiLSTM층은 513개의 출력 노드 수를 가지며 두 번째 BiLSTM층의 출력 노드 수는 1024개이다. 이후, BiLSTM의 출력은 첫 번째와 두 번째 전결합층을 통과하며 각각의 전결합층의 출력 노드 수는 513개이다.

이후, 신경망 출력이 sigmoid 활성화 함수까지 도달하게 되면 음성에 대한 다채널 이진 마스크를 생성한다. 그림 2에서 보는 바와 같이, 다채널 음성에 대한 이진 마스크는 median filter를 통과하게 된다. Median filter를 통과한 후에는 단일 채널 음성 이진

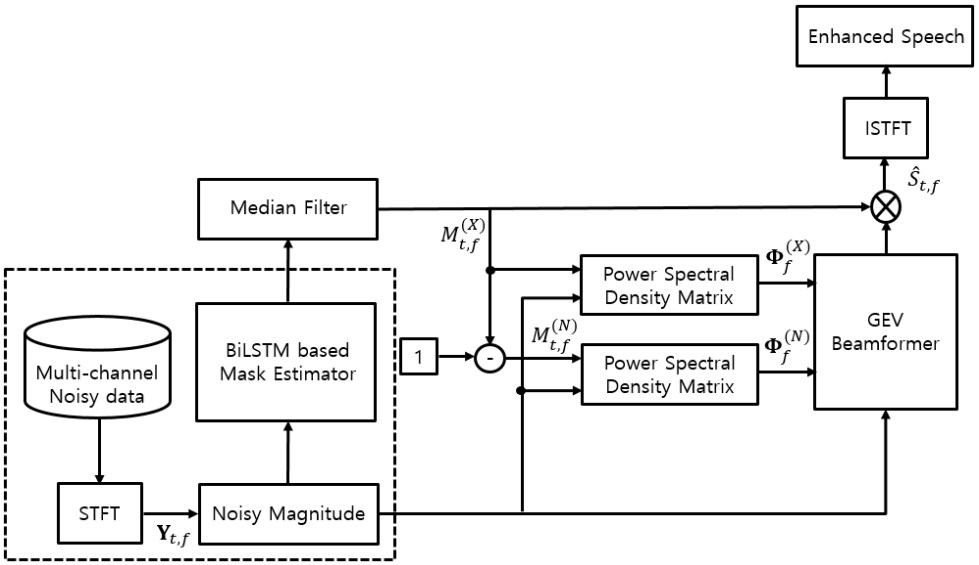

그림 2. GEV 빔형성을 위한 제안된 BiLSTM 기반 이진 마스크 추정 기법 구성도
Fig. 2. Block diagram of the proposed BiLSTM-based binary mask estimation method for GEV beamforming

마스크 $M_{t,f}^{(X)}$로 생성되며 잡음에 대한 이진 마스크는 $(1-M_{t,f}^{(X)})$ 연산으로 생성된다. 또한, 음성과 잡음에 대한 이진 마스크가 구해짐에 따라 음성 및 잡음에 대한 전력 스펙트럼 밀도행렬을 계산할 수 있으며 식 (5)-(8)에 의하여 GEV 빔형성 가중치를 얻을 수 있다.

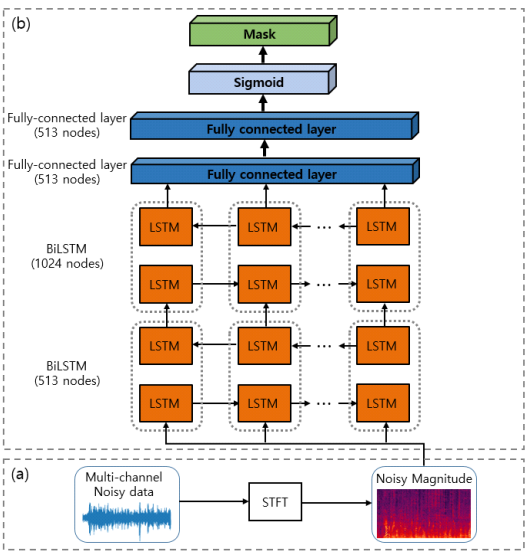

그림 3. BiLSTM 신경망 기반 이진 마스크 추정 모델; (a) 특징 추출 과정, (b) 신경망 구조
Fig. 3. BiLSTM-based binary mask estimation model; (a) a feature extraction process and (b) a neural network architecture

이후에는 다채널 입력 신호와 빔형성 가중치의 가중합으로 나타낼 수 있으며 빔형성 출력단계에서 후처리 과정으로 음성에 대한 이진 마스크를 적용할 수 있다.

앞에서 기술한 바와 같이, 이진 마스크를 빔형성 가중치를 구하는 과정과 빔형성 출력단에 적용함으로써 잡음 성분을 최소화할 수 있으며 이로 인한 향상된 음성에 대한 추정치 $\hat{S}_{f,t}$는 다음과 같다.

$$\hat{S}_{t,f} = \mathbf{w}_f^H \mathbf{Y}_{t,f} \cdot M_{t,f}^{(X)} \tag{12}$$

마지막으로, 향상된 음성을 추정한 다음에는 역 단시간 푸리에 변환(inverse short-time Fourier transform, ISTFT)을 통해 신호를 주파수 영역에서 시간 영역으로 변환하며 최종적으로 향상된 단일 채널 음성 신호를 얻을 수 있다.

Ⅳ. 실험 결과

본 절에서는 제안된 방법의 성능을 평가하기 위해 다음과 같이 환경 구축 및 실험을 진행하였다.

4.1 실험 환경

PC 환경은 Intel(R) Core(TM) i7-7700 CPU 3.60GHz, 64GB RAM 환경에서 실험을 진행하였다. 신경망 학습 도구로는 Python 기반 Pytorch 프레임워크를 활용하였으며 GPU 모델은 Nvidia GeForce

GTX 1080Ti를 사용하였다.

4.2 평가 데이터셋

제안된 방법의 실험을 위해 CHiME-3 데이터셋을 구축하여 평가를 진행하였다. CHiME-3 데이터셋은 16 kHz로 샘플링(sampling)된 음원과 잡음이 많은 환경을 기반으로 하며 잡음 환경은 bus, cafe, pedestrian area, street로 구성되어 있다. 발화자의 음성은 6채널로 녹음되어 있으며 7,138개 발화 문장으로 구성된 모의 훈련 데이터와 1,320개 발화 문장의 테스트 데이터로 이루어져 있다.

4.3 평가 방법

본 논문에서는 기존의 GEV 빔형성 모델[17]과 제안한 모델을 비교한다. 두 모델 모두 동일한 CHiME-3 데이터셋의 6채널 음원을 데이터로 사용하였으며 bus, cafe, pedestrian area, street 총 4가지 잡음 환경에서의 훈련 데이터로 학습을 진행하였다. 또한, GEV 빔형성 성능을 평가하기 위해 객관적 음질평가 방법으로 perceptual evaluation of speech quality (PESQ)[18] 점수와 신호 왜곡도(signal-to-distortion ratio, SDR)[19] 수치를 사용하였다.

PESQ는 음성품질을 객관적으로 평가하기 위해 개발된 평가 수치이다. PESQ는 기준이 되는 원 음성 신호와 평가하려는 음성과 비교하여 명료도 면에서 원 음성 신호와 얼마나 유사한지 평가하여 mean opinion score (MOS)로 나타내며, -0.5에서 4.5 사이의 값을 가지며 원음과 유사할수록 4.5에 가까운 값을 나타낸다.

SDR은 음원의 잡음 외에 잔향과 시스템 잡음을 모두 포함하여 종합적인 성능을 평가할 수 있는 지수이다. 먼저 추정된 음성 신호를 다음과 같이 표현할 수 있다.

$$\hat{S}_{\text{target}} = S_{\text{target}} + e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}} \tag{13}$$

여기서, S_{target}은 실제 음성을, $\hat{S}_{\text{target}}$은 추정된 음성을 의미하며 e_{noise}는 음원에 섞인 잡음, e_{interf}는 주변 배경 잡음이나 반향(reverberation) 신호를 나타내는 간섭 잡음(interference noise)이다. 또한 e_{artif}는 인공 결함 잡음(artifact noise)을 의미하며 녹음기기 또는 마이크 자체의 시스템 잡음이 이에 해당한다. 실제 신호 대비 추정된 신호의 전체적인 오차의 에너지 비를 나타내는 SDR은 다음과 같이 정의한다.

$$\text{SDR} = 10\log_{10}\frac{\| S_{\text{target}} \|^2}{\| e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}} \|^2}\,(\text{dB}) \tag{14}$$

4.4 성능 평가

그림 4는 CHiME-3 데이터셋 평가 음원에 이진 마스크 추정 기반의 GEV 빔형성을 적용하여 얻은 개선된 음성의 스펙트로그램을 비교하여 보여준다. 그림 4(a)에서는 전체적으로 음성에 잡음을 포함하고 있는 스펙트럼을 볼 수 있다. 특히, 저주파 대역에 잡음이 많이 존재하고 있다. 그림 4(b)는 기존의 BiLSTM 기반 GEV 빔형성[17]을 적용한 결과이며, 그림 4(a)와 비교해 볼 때, 잡음이 많이 제거된 것을 볼 수 있지만 아직 여전히 잡음 성분이 스펙트럼에서 남아있는 것을

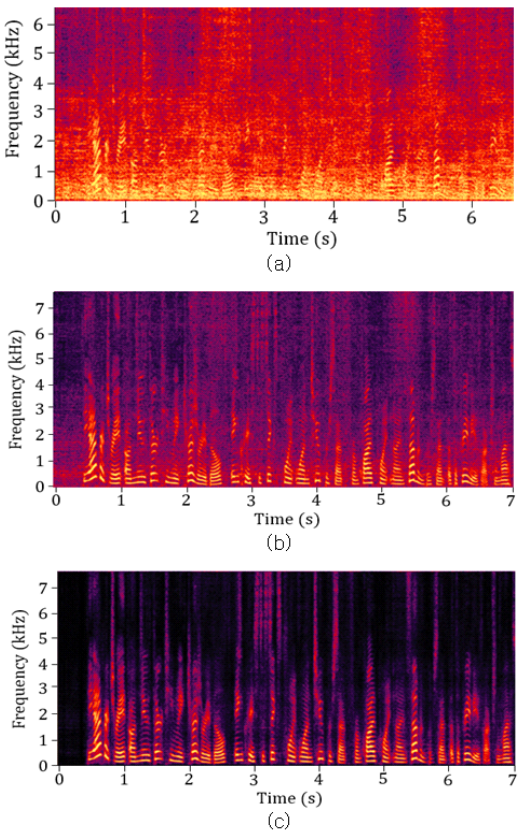

그림 4. 스펙트로그램 비교; (a) 잡음 음성, (b) 기존의 BiLSTM 기반 마스크 추정 기법을 적용한 GEV 빔포밍으로 개선된 음성, (c) 제안된 BiLSTM 기반 마스크 추정 후처리 기법을 적용한 GEV 빔포밍으로 개선된 음성

Fig. 4. Spectrogram comparison; (a) noisy input speech, (b) enhanced speech by GEV beamforming employing a conventional BiLSTM-based mask estimation, and (c) enhanced speech by GEV beamforming employing the proposed post-processing method of BiLSTM-based mask estimation with post-IBM

볼 수 있다. 반면, 그림 4(c)는 본 논문에서 제안된 방법을 적용한 스펙트로그램으로 그림 4(b)와 비교했을 때 상대적으로 잡음이 더 제거된 것을 볼 수 있다.

표 1은 기존 방법들과 비교하여 객관적 음질평가 결과를 보여준다. 표에서 보는 바와 같이, 기존 GEV 빔형성 모델은 CHiME-3 데이터의 bus, cafe, pedestrian area, street 모든 잡음 환경에 대해서 PESQ, SDR의 평균을 취했을 때 PESQ는 2.71 MOS 이며 SDR은 6.67 dB인 것을 확인할 수 있다. 그리고 제안한 방식의 PESQ는 3.05 MOS이며 SDR은 7.58 dB로 기존 GEV 빔형성 보다 PESQ 점수와 SDR 수치에서 각각 0.34 MOS와 0.91 dB를 개선할 수 있었다.

기존 BiLSTM 기반 GEV 빔형성과 달리 제안된 방법은 이진 마스크를 GEV 가중치를 구하는 과정과 빔형성 출력에 적용하여 GEV 빔형성 전체적인 과정에 영향을 준다. 또한, 기존 GEV 빔형성은 음성 및 잡음에 대한 이진 마스크를 BiLSTM 신경망의 타겟으로 하나, 제안된 방법은 음성에 대한 이진 마스크만을 타겟으로 훈련하여 계산량에서 장점을 지닌다. BiLSTM 신경망 구조에서는 BiLSTM 2층을 쌓고 층마다 배치 정규화를 적용하여 보다 깊은 신경망을 훈련하게 하였다. 이러한 구조로 인해 표 1에서 볼 수 있듯이, 기존 GEV 빔형성 보다 높은 성능을 보여주는 것을 알 수 있다.

표 1. 기존 및 제안된 BiLSTM 기반 이진 마스크 추정 방법을 갖는 GEV 빔형성의 PESQ와 SDR 비교
Table 1. Comparison of PESQ and SDR of GEV beamforming employing a conventional and proposed BiLSTM-based binary mask estimation method

Noise Type	Model	PESQ	SDR (dB)
Bus	Noisy data (6-ch)	1.71	0.28
	BiLSTM[17]	2.88	5.70
	BiLSTM+Post-IBM	3.14	8.14
Cafe	Noisy data (6-ch)	1.51	1.15
	BiLSTM[17]	2.60	7.01
	BiLSTM+Post-IBM	3.01	7.20
Pedestrian	Noisy data (6-ch)	1.50	1.26
	BiLSTM[17]	2.68	7.03
	BiLSTM+Post-IBM	3.04	7.14
Street	Noisy data (6-ch)	1.51	0.63
	BiLSTM[17]	2.69	6.92
	BiLSTM+Post-IBM	3.02	7.83
Avg.	Noisy data (6-ch)	1.56	0.83
	BiLSTM[17]	2.71	6.67
	BiLSTM+Post-IBM	3.05	7.58

V. 결 론

본 논문에서는 GEV 빔형성을 위한 BiLSTM 기반 이진 마스크 후처리 기법을 제안하였다. 제안된 방법은 잡음 환경에서 BiLSTM 신경망을 통해 다채널 잡음에 대한 음성을 입력으로 하여 clean 음성에 대한 이진 마스크를 목표로 학습하였다. 이후, BiLSTM으로 추정된 이진 마스크를 적용하여 음성 및 잡음의 전력 스펙트럼 밀도 행렬을 구하고, 이에 대한 고유값 분해식으로 GEV 가중치를 계산하였다. 이처럼 음성 및 잡음에 대한 이진 마스크에 의해 전력 스펙트럼 밀도가 구해지며 이에 따라 빔형성의 가중치가 결정되기 때문에 GEV 빔형성의 성능을 개선하는 데 있어서 이진 마스크의 역할이 중요한 것을 알 수 있었다. 또한, GEV 빔포밍 처리된 음성의 음질향상을 위해 빔형성 출력단계에서는 음성에 대한 이진 마스크를 적용하였다. 제안된 방법을 CHiME-3 데이터셋으로 성능 평가한 결과, 기존 GEV 빔형성 기법의 PESQ는 2.71 MOS, SDR은 6.67 dB이며 제안된 방법의 PESQ는 3.05 MOS, SDR은 7.58 dB로 기존 GEV 빔형성 기법보다 PESQ와 SDR 수치에서 각각 0.34 MOS와 0.91 dB를 개선하였다.

References

[1] S. H. Kim and J. Y. Ahn, "Speech recognition system in car noise environment," *J. Digital Contents Soc.*, vol. 10, no. 1, pp. 121-127, Mar. 2009.

[2] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Berlin, Germany, Springer-Verlag, 2008.

[3] H. J. Kim and H. S. Yoon, "Multi-directional simultaneous speech recognition based on beamforming," in *Proc. KIISE KCC,* vol. 37, no. 1, pp. 179-183, Jeju Island, Korea, Jun. 2010.

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

[5] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized

eigenvalue decomposition," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 5, pp. 1529-1539, Jul. 2007.

[6] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *Proc. IEEE ICASSP*, pp. 121-124, Honolulu, HI, Apr. 2007.

[7] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: dataset, task and baselines," in *Proc. IEEE ASRU*, pp. 504-511, Scottsdale, AZ, Dec. 2015.

[8] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *Proc. IEEE Workshop ASRU*, pp. 444-451, Scottsdale, AZ, Dec. 2015.

[9] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673-2681, Nov. 1997.

[10] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Sign. Process.*, vol. 50, no. 9, pp. 2230-2244, Sep. 2002.

[11] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, Springer, 2005.

[12] Y. Li and D. Wang, "On the optimality of ideal binary time-frequency masks," in *Proc. IEEE ICASSP*, pp. 3501-3504, Las Vegas, NV, Mar. 2008.

[13] G. W. Lee and H. K. Kim, "Multi-task learning U-net for single-channel speech enhancement and mask-based voice activity detection," *Applied Sci.*, vol. 10, no. 9, May 2020.

[14] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artificial Intell. and Statistics*, pp. 249-256, Sardinia, Italy, Mar. 2010.

[15] S. Loffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, pp. 448-456, Lille, France, Jul. 2015.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929-1958, Jan. 2014.

[17] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE ICASSP*, pp. 196-200, Shanghai, China, Mar. 2016.

[18] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 16, no. 1, pp. 229-238, Jan. 2008.

[19] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 4, pp. 1462-1469, Jul. 2006.

송 일 훈 (Ilhoon Song)

2019년 2월 : 한국산업기술대학교 전자공학부 학사

2020년 3월~현재 : 광주과학기술원 전기전자컴퓨터공학부 석사과정

<관심분야> 음성 방향 추정, 동영상 이상 상황 검출

[ORCID:0000-0003-1768-3103]

김 홍 국 (Hong Kook Kim)

1988년 2월 : 서울대학교 제어계측공학과 공학사
1990년 2월 : 한국과학기술원 전기 및 전자공학과 공학석사
1994년 8월 : 한국과학기술원 전기 및 전자공학과 공학박사
1990년~1998년 : 삼성종합기술원 전문연구원
1998년~1998년 : MMC Technology 선임연구원
1998년~2003년 : AT&T Labs-Research Senior Member Technical Staff
2014년~2015년 : City University of New York, Visiting Professor
2003년 8월~현재 : 광주과학기술원 전기전자컴퓨터공학부 및 AI 대학원 교수
<관심분야> 음성 · 오디오 처리, 음성인식, 음향 기반 헬스케어, 딥러닝 기반 시계열 예측
[ORCID:0000-0002-0105-6693]