한국음향학회지 제40권 제5호 pp. 439~451 (2021)
The Journal of the Acoustical Society of Korea Vol.40, No.5 (2021)
https://doi.org/10.7776/ASK.2021.40.5.439

pISSN : 1225-4428
eISSN : 2287-3775

잡음 환경에서의 음성인식을 위한 온라인 빔포밍과 스펙트럼 감산의 결합

Combining deep learning-based online beamforming with spectral subtraction for speech recognition in noisy environments

윤성욱,[1] 권오욱[1,2†]

(Sung-Wook Yoon[1] and Oh-Wook Kwon[1,2†])

[1]충북대학교 지능로봇공학과, [2]충북대학교 컴퓨터정보통신연구소

(Received June 10, 2021; accepted August 9, 2021)

초 록: 본 논문에서는 실제 환경에서의 연속 음성 강화를 위한 딥러닝 기반 온라인 빔포밍 알고리듬과 스펙트럼 감산을 결합한 빔포머를 제안한다. 기존 빔포밍 시스템은 컴퓨터에서 음성과 잡음을 완전히 겹친 방식으로 혼합하여 생성된 사전 분할 오디오 신호를 사용하여 대부분 평가되었다. 하지만 실제 환경에서는 시간 축으로 음성 발화가 띄엄띄엄 발성되기 때문에, 음성이 없는 잡음 신호가 시스템에 입력되면 기존 빔포밍 알고리듬의 성능이 저하된다. 이러한 효과를 경감하기 위하여, 심층 학습 기반 온라인 빔포밍 알고리듬과 스펙트럼 감산을 결합하였다. 잡음 환경에서 온라인 빔포밍 알고리듬을 평가하기 위해 연속 음성 강화 세트를 구성하였다. 평가 세트는 CHiME3 평가 세트에서 추출한 음성 발화와 CHiME3 배경 잡음 및 MUSDB에서 추출한 연속 재생되는 배경음악을 혼합하여 구성되었다. 음성인식기로는 Kaldi 기반 툴킷 및 구글 웹 음성인식기를 사용하였다. 제안한 온라인 빔포밍 알고리듬 과 스펙트럼 감산이 베이스라인 빔포밍 알고리듬에 비해 성능 향상을 보임을 확인하였다.
핵심용어: 온라인 빔포밍, 딥 러닝, 스펙트럼 감산, 연속 음성 강화

ABSTRACT: We propose a deep learning-based beamformer combined with spectral subtraction for continuous speech recognition operating in noisy environments. Conventional beamforming systems were mostly evaluated by using pre-segmented audio signals which were typically generated by mixing speech and noise continuously on a computer. However, since speech utterances are sparsely uttered along the time axis in real environments, conventional beamforming systems degrade in case when noise-only signals without speech are input. To alleviate this drawback, we combine online beamforming algorithm and spectral subtraction. We construct a Continuous Speech Enhancement (CSE) evaluation set to evaluate the online beamforming algorithm in noisy environments. The evaluation set is built by mixing sparsely-occurring speech utterances of the CHiME3 evaluation set and continuously-played CHiME3 background noise and background music of MUSDB . Using a Kaldi-based toolkit and Google web speech recognizer as a speech recognition back-end, we confirm that the proposed online beamforming algorithm with spectral subtraction shows better performance than the baseline online algorithm.
Keywords: Online beamforming, Deep learning, Spectral subtraction, Continuous speech enhancement

PACS numbers: 43.60.Fg, 43.72.Dv

† **Corresponding author:** Oh-Wook Kwon (owkwon@cbnu.ac.kr)
Department of Intelligent Systems and Robotics, Chungbuk National University, 1 Chungdae-ro, Seowon-gu, Cheongju, Chungcheongbuk-do 28644, Republic of Korea
(Tel: 82-43-261-3374, Fax: 82-43-268-2386)

I. 서 론

스펙트럼과 공간 정보를 활용한 다채널 음성 향상은 자동 음성인식(Automatic Speech Recognition, ASR)의 성능 향상에 효과적인 방법임이 입증되었다.[1,2] 전통적인 다채널 음성 향상 방법으로 다채널 음수 미포함 행렬 분해(Multichannel Nonnegative Matrix Factorization, MNMF),[3] 다채널 위너 필터(Multichannel Wiener Filter, MWF), [4] 빔포밍이 ASR 성능 향상을 위한 주요 기술로 사용되었다. 이중 빔포밍은 성능 향상에 가장 중요한 기술로써 최소 분산 무왜곡 응답(Minimum-Variance Distortionless Response, MVDR),[5] 일반화 고유값(Generalized Eigen Value, GEV),[6] 일반 부엽 제거기(Generalized Sidelobe Canceller, GSC)[7] 등이 활발하게 연구되었다.

최근 심층신경망(Deep Neural Network, DNN)이 ASR에서 주목할 만한 성능 향상을 보였다. 딥러닝을 이용한 최신의 빔포밍 기술로 시간-주파수(Time-Frequency, T-F) 마스크 추정 방법이 제안되었다. 많은 연구에서 딥러닝 기반 마스크 추정 빔포밍이 성공적으로 적용되었고,[8-10] 전통적인 빔포밍의 성능을 뛰어넘었다. 딥러닝이 성공적으로 빔포밍에 적용됨에 따라 실제 환경에서도 어느 정도 안정적인 성능을 보였다.

이러한 발전에도 불구하고, 기존 빔포밍 알고리듬 및 평가 세트는 실제 환경에서 적용되기에는 많은 부분이 아직도 충분히 고려되지 못하였다. 대부분의 기존 빔포밍 알고리듬은 음성과 잡음이 완전히 겹쳐진 사전 분절된 발화 단위 잡음 음성을 대상으로 주로 연구되었다. 실제 사용 환경을 고려하면 잡음은 항상 존재하지만, 사용자 음성이 희박한 연속 잡음 음성이 빔포머의 입력이 된다. 연속된 잡음 음성 처리를 위해 기존 시불변 빔포밍 벡터가 아닌 잡음 음성에 적응할 수 있는 시변 빔포밍 벡터 추정을 위한 온라인 빔포밍 알고리듬이 필요하다. 또한 실제 사용 시 시간 영역에서 음성 구간이 희박하고, 잡음만이 존재하는 프레임 입력은 온라인 빔포밍 알고리듬의 성능 열화로 이어지기 때문에 이를 위한 대책이 필요하다.

기존 연구에서는 음성 활동 감지기(Voice Activity Detector, VAD)의 프레임 단위 출력인 음성/잡음 레이블정보를 이용하여 잡음만이 존재하는 프레임을 빔포밍에 사용하지 않거나,[11,12] 학습 데이터에서 주어진 공간 사전 정보를 이용하여 잡음만이 존재하는 입력으로부터 음성신호에 대한 전력 스펙트럼 밀도(Power Spectral Density, PSD) 행렬의 과적합을 방지하여 성능 향상을 달성하였다.[13] 그러나 낮은 신호대잡음비(Signal to Noise Ratio, SNR) 환경에서 VAD 기반 빔포밍의 성능은 크게 열화 되며, 학습 데이터에서 주어진 공간 사전 정보를 이용한 방법은 학습 데이터의 화자 및 공간적 배열에 의존하기 때문에 학습과 테스트 환경의 차이에 취약하다.

본 논문에서는 낮은 신호대잡음비의 실제 환경에서 강인한 온라인 빔포밍 시스템을 위해 공간 사전 정보에 영향을 받지 않는 딥러닝 기반 마스크 추정 빔포머를 이용한 온라인 빔포밍 알고리듬을 제안하고 추가적으로 스펙트럼 감산을 결합한 시스템으로 성능 향상을 거두었다. 기존의 온라인 빔포밍 알고리듬의 성능평가를 위한 평가 세트는 음성과 잡음이 완전히 겹쳐진 사전 분절된 발화 단위 음성 평가 세트에서 진행되었다. 긴 시간의 연속 잡음 음성을 평가 세트로 사용하는 경우에도 대부분의 시간 영역에서 음성이 존재하기 때문에 잡음만 존재하는 구간에서 온라인 빔포밍 알고리듬의 성능 열화가 되는 현상을 고려하지 않은 평가가 진행되었다. 실제 환경을 고려한 평가를 위해, 본 논문에서는 CHiME3[14]와 MUSDB[15]를 활용하여 음성 구간이 희박한 연속된 잡음 음성 스트림으로 이루어진 Continuous Speech Enhancement (CSE) 평가 세트를 제작하였다.

2절에서는 Bidirectional Long Short-Term Memory (BLSTM) 기반 마스크 추정 빔포밍에 관해 설명하고, 3절에서는 제안하는 온라인 업데이트 알고리듬 및 스펙트럼 감산 빔포밍 시스템에 대해 설명한다. 4절에서는 CSE 평가 세트를 이용해 제안 시스템의 성능 평가를 진행하고, 제안 알고리듬의 성능을 검증한다. 5절에서 결론을 맺는다.

II. BLSTM 기반 마스크 추정 빔포밍

딥러닝을 이용한 마스크 추정 기반 빔포밍은 기존 빔포밍 알고리듬에서 중요한 PSD 행렬의 추정을 딥러

닝으로 대체하여 획기적인 성능 향상을 이루었다[8]. 기존의 빔포밍 방식이 공간 배열에 의존적인데 반해 딥러닝을 이용한 마스크 추정 빔포밍은 마이크의 개수, 배열과 관계없이 독립적으로 이루어진다. 본 논문에서는 CHiME3 챌린지[14]에서 우수한 성능을 보인 BLSTM 기반 마스크 추정 빔포밍[8]을 이용한다.

2.1 BLSTM 기반 마스크 추정 빔포밍

스펙트럼 기반 마스크 추정은 BLSTM이 사용된다. 각 채널은 동일한 BLSTM 가중치를 공유하며, 각 채널에서 음성 마스크 $M_X(t,f)$와 잡음 마스크 $M_N(t,f)$가 독립적으로 추정된다. BLSTM의 학습 타겟으로는 이상적인 이진 마스크(Ideal Binary Mask, IBM)가 사용되었다.

채널 개수를 C로 정의한다. C개의 음성 마스크와 C개의 잡음 마스크를 추정할 수 있다. 추정된 C개의 마스크를 하나로 합치는 풀링이 필요한데 여러 가지 풀링 옵션 중 중앙값을 사용하였다. 이것은 불특정 채널에서 마스크 추정에 실패하여 특이 값이 발생할 경우에도 안정된 값을 내어 줄 수 있기 때문이다. 중앙값을 통해 합쳐진 음성 마스크와 잡음 마스크를 이용하여 최종적으로 음성과 잡음에 관한 PSD 행렬이 추정된다.

2.1.1 마스크 추정 네트워크 구성

마스크 추정을 위한 신경망은 4층으로 구성된다. Fig. 1 및 Table 1에서 전체적인 구조를 확인할 수 있다. 잡음 음성 스트림을 16 kHz 샘플링 후 1,024 프레임 사이즈, 256 프레임 쉬프트 사이즈를 사용해 Short Time Fourier Transform(STFT)를 한다. STFT의 결과로부터 513개의 스펙트럼 크기를 취하여 마스크 추정을 위한 신경망의 입력으로 사용한다.

첫 번째 층은 256 출력 유닛 BLSTM 층으로 이루어져 있으며 tanh 을 활성화 함수로 사용한다. BLSTM의 메모리 cell 개수는 1,024개이다. 다음 2,3층은 513 유닛을 갖는 순방향(Feed Forward, FF) 층으로 이루어져 있으며, 정류 선형 유닛(Rectified Linear Unit, ReLU)을 활성화 함수로 사용한다. 입력의 513-포인트 스펙트럼의 크기를 고려하여 513 유닛을 사용하였다. 입력의 마지막 층은 1026 유닛으로 구성되며 2개의 부분으로 나누어지며, 1 ~ 513 유닛은 $M_X(t,f)$를 추정하고, 514 ~ 1,026 유닛은 $M_N(t,f)$을 추정한다. 활성화 함수로는 sigmoid를 사용하여, 0 ~ 1 사이의 값을 추정한다. 각 시간-주파수 빈(bin)에서 음성, 잡음 마스크 추정 값의 합이 1이 되는 제약을 두지 않았다.[8]

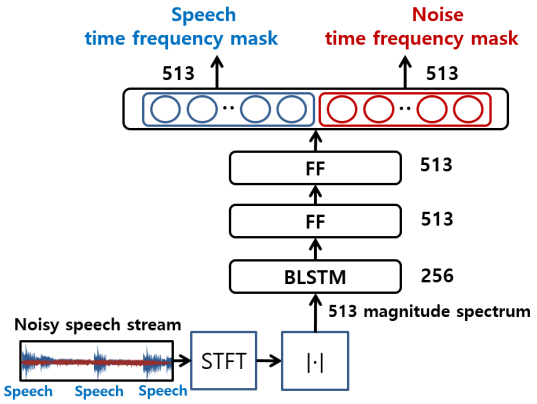

Fig. 1. (Color available online) Neural network for mask estimation.

Table 1. Neural network architecture for mask estimation.

	Units	Type	Activation	Dropout
Layer 1	256	BLSTM	tanh	0.5
Layer 2	513	FF	ReLU	0.5
Layer 3	513	FF	ReLU	0.5
Layer 4	1026	FF	sigmoid	0.0

2.1.2 정답 마스크

IBM이 학습 마스크의 정답으로 사용되었다. Eqs. (1), (2)에서 $|X(t,f)|$는 목표 음성의 스펙트럼 크기, $|N(t,f)|$는 잡음 스펙트럼 크기이다. 여기서 t는 프레임 인덱스, f는 주파수 빈의 인덱스이다. IBM_X는 목표 음성의 정답 마스크 값, IBM_N은 잡음의 정답 마스크 값이다.

$$IBM_X(t,f)=\begin{cases}1, & \dfrac{|X(t,f)|^2}{|N(t,f)|^2}>10^{th_X(f)/10} \; and \; |X(t,f)|^2>0.005\times 10^{th_X(f)/10}\\ 0, & \text{otherwise}\end{cases},$$
$$th_X(f)=\ th\_voiced\_speech*voiced(f)+ th\_unvioced\_speech*unvoiced(f)$$
$$th\_voiced\_speech=5,\ th\_unvioced\_speech=0 \quad (1)$$

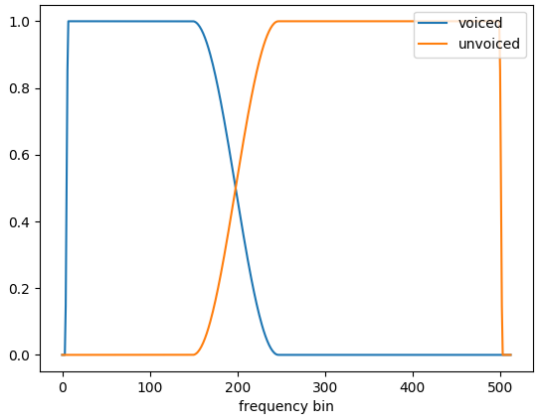

Fig. 2. (Color available online) Graph of voiced(f) and unvoiced(f).

$$IBM_N(t,f)=\begin{cases}1, & \dfrac{|X(t,f)|^2}{|N(t,f)|^2}<10^{th_N(f)/10}\ or\ |X(t,f)|^2<0.005\times 10^{th_N(f)/10}\\ 0, & \text{otherwise}\end{cases},$$
$$th_N(f)=th\_unvoiced\_noise*voiced(f)+th\_voiced\_noise*unvoiced(f)$$
$$th\_unvoiced\_noise=-10,\ th\_voiced\_noise=-10 \quad (2)$$

두 개의 문턱 값 $th_X(f)$와 $th_N(f)$은 동일하지 않다. 문턱 값에 사용된 *voiced(f)*와 *unvoiced(f)*는 주파수 빈의 개수에 따라 결정되며, Fig. 2와 같이 0 ~ 1사이의 값을 갖는다. 대부분의 음성 에너지가 1 kHz 이하에 분포해 있으며, 고주파로 갈수록 작아지는 특성을 모델링하였고 음성이 아닌 신호에 대한 특성 또한 고려한 모델링을 하였다. $th_X(f)$는 음성의 정보가 많은 저주파에서는 신호대잡음비가 5 dB 이상일 때 음성 클래스로 결정되게 하며, 고주파에서는 신호대잡음비 0 dB 이상에서 음성 클래스로 결정되게 한다. 동시에 목표 음성 스펙트럼의 전력이 충분히 크면서 신호대잡음비 조건을 만족할 때 최종 음성 클래스로 결정한다. $th_N(f)$는 주파수 빈 전역에 걸쳐 신호대잡음비가 −10 dB 이하에서만 잡음 클래스로 결정되게 한다. 또한, 목표 음성 신호의 전력을 고려하여, 음성이 매우 미미한 경우에도 잡음 클래스로 결정하게 된다. 낮은 오인식률(False Acceptance Rate, FAR)을 보장하기 위해 프레임 SNR이 충분히 높거나 낮은 경우에만 음성이나 잡음 클래스로 결정한다. 동시에 목표 음성의 스펙트럼 크기를 고려해 음성의 크기가 충분히 작으면서 SNR 조건을 만족할 때 최종 잡음 클래스로 결정한다. IBM 마스크는 학습 및 개발 세트의 데이터에서만 계산되어 학습에 사용된다.

2.1.3 손실 함수

신경망의 학습은 이진 분류에 사용되는 이진 크로스 엔트로피를 손실 함수로 사용하며 Eq. (3)과 같이 표현된다.

$$L=-\frac{1}{F}\frac{1}{2T}\sum_{v\in(N,X)}\sum_{t=1}^{T}\sum_{f=1}^{F}IBM_v(t,f)\log M_v(t,f)+(1-IBM_v(t,f))\log(1-M_v(t,f)),\ v\in\{X,N\}, \quad (3)$$

여기서 t는 프레임 인덱스, T는 한 발화의 프레임 길이, f는 주파수 빈의 인덱스, F는 총 주파수 빈 수를 나타낸다. $M_V(t,f)$는 신경망 4층에서 추정된 마스크 값으로서 1 ~ 513 유닛에서는 음성 마스크 값 $M_X(t,f)$이 추정되고 514 ~ 1026 유닛에서는 잡음 마스크 값 $M_N(t,f)$이 동시에 추정된다.

2.2 음향 빔포밍

자주 사용되는 음향 빔포밍인 MVDR과 GEV에 대해 기술한다. 두 빔포밍 방식 모두 Eq. (4)와 같이 추정된 마스크 M_X와 M_N 관측 신호의 스펙트럼 $Y(t,f)$로부터 PSD 행렬을 추정한다.

$$\Phi_{vv}(f)=\sum_{t=1}^{T}M_v(t,f)Y(t,f)Y(t,f)^H,\ v\in\{X,N\}. \quad (4)$$

H는 에르미트 연산자이며, PSD 행렬의 차원은 $F\times C\times C$으로 다채널 마이크 사이의 음성과 잡음의 전력 분포를 의미한다.

2.2.1 MVDR 빔포머

음성인식에서 가장 자주 사용되는 빔포밍 방식은 MVDR 빔포머이다. MVDR 빔포머의 수식은 Eq. (5)와 같이 계산된다. 목표 음성 신호의 방향을 유지하는 제약 조건을 가지고 잔여 잡음을 최소화하는 W

를 구하면 MVDR 빔포밍 벡터 W_{MVDR}이 된다. 반응 벡터 d는 도착각 추정을 통해 얻을 수 있으며 대체적인 방법으로는 음성 PSD 행렬 고유 값 분해를 이용하면 얻을 수 있다.[16] 마스크 추정 기반 MVDR 빔포밍에서는 음성 PSD 행렬을 고유 값 분해하여 d를 추정한다.

$$W_{MVDR} = \arg\min_{W} W\Phi_{NN}W \quad \text{subject to} \quad W^{H}d = 1. \tag{5}$$

라그랑주 승수법을 이용하면 W_{MVDR}는 Eq. (6)과 같이 표현 가능하다.

$$W_{MVDR} = \frac{\Phi_{NN}^{-1}d}{d^{H}\Phi_{NN}^{-1}d}. \tag{6}$$

2.2.2 GEV 빔포머

GEV 빔포밍 벡터 W_{GEV}는 Eq. (7)과 같이 Rayleigh coefficient에서 주파수 빈별 신호대잡음비를 최대화함으로써 구할 수 있다.[17] 최적화 해답은 Eq. (8)로 주어진다.

$$W_{GEV} = \arg\min_{W} \frac{W^{H}\Phi_{XX}W}{W^{H}\Phi_{NN}W}. \tag{7}$$

$$W_{GEV} = P(\Phi_{NN}^{-1}\Phi_{XX}). \tag{8}$$

$P(\bullet)$는 가장 큰 고유값에 대응되는 고유벡터를 구하는 함수를 의미한다. 추정을 위해 음성 소스와 마이크 배열 사이에 전달 함수 및 잡음의 공간적 정보는 전혀 필요하지 않다.

본 연구에서 MVDR 빔포머 대신 GEV 빔포머를 사용하는 이유는 Eq. (6)에서와 같이 MVDR 빔포밍 벡터는 항상 잡음 PSD의 역행렬 계산이 필요하기 때문에 특정 주파수 빈 값이 부족할 경우 수학적으로 매우 불안정해지고, 성능 열화로 이어진다. 반면에 GEV빔포밍 벡터 계산 Eq. (8)에 슈어 분해를 이용하면 역행렬 연산을 피하는 동시에 수학적 안정성을 보장할 수 있으며 CHiME3 챌린지에서 MVDR 빔포밍에 비해 좋은 성능을 보임이 확인되었다.[16]

III. 제안 구조

본 장에서는 연속 잡음 음성 처리를 위해 BLSTM 마스크 추정 값을 이용한 온라인 빔포밍 알고리듬 시스템을 제안한다. 연속 잡음 음성 처리를 위해 잡음환경에 적응하는 PSD 행렬 업데이트 알고리듬이 필요하다. 본 연구에서는 블록 단위 PSD 행렬 업데이트를 제안한다. 또한, 낮은 신호대잡음비 환경에서 추가적인 성능 향상을 위하여 스펙트럼 감산을 결합한 빔포밍 시스템을 제안한다. 마지막으로 PSD 행렬의 빠른 수렴을 위해 블록을 나누어 PSD 행렬을 업데이트 하는 알고리듬을 제안한다.

3.1 PSD 행렬 업데이트 알고리듬

연속된 잡음 음성 처리를 위해, 기존 발화단위 PSD 행렬 추정 Eq. (4)를 연속해 들어오는 관측 신호에 적합하게 업데이트해 주어야 한다. 이를 위해 BLSTM 마스크 추정 값을 이용한 시변 PSD 행렬 추정 알고리듬을 제안하고, PSD 행렬 추정시 통계적 충분성을 보장하기 위해 링 버퍼를 사용하였다.

3.1.1 마스크 추정 값을 이용한 PSD 행렬 업데이트 알고리듬

본 연구에서는 블록 단위 업데이트 알고리듬을 제안한다. l은 블록 인덱스, L은 한 블록 안에 포함된 프레임 수이다. l번째 블록에서 추정된 PSD 행렬 $\Phi_{vv}^{l}(f)$는 Eq. (9)와 같이 계산된다.

$$\Phi_{vv}^{l}(f) = \sum_{t=1}^{L} M_{v}(t,f)Y(t,f)Y(t,f)^{H}. \tag{9}$$

추정된 $\Phi_{vv}^{l}(f)$는 Eq. (10)과 같이 가중치 $\alpha_{V}^{l}(f)$를 이용해 누적 추정된 $\Phi_{vv}^{l-1}(f)$와 가중 합산되고 l 번째 블록에서 누적 추정된 PSD $\Phi_{vv}^{l}(f)$가 얻어진다.

$$\Phi_{vv}^{l}(f) = \begin{cases} \sum_{t=1}^{L} M_{v}(t,f)Y(t,f)Y(t,f)^{H}, & l = 1 \\ \alpha_{v}^{l}(f)\hat{\Phi}_{vv}^{l}(f) + (1-\alpha_{v}^{l}(f))\Phi_{vv}^{l-1}(f), & \text{otherwise} \end{cases}. \tag{10}$$

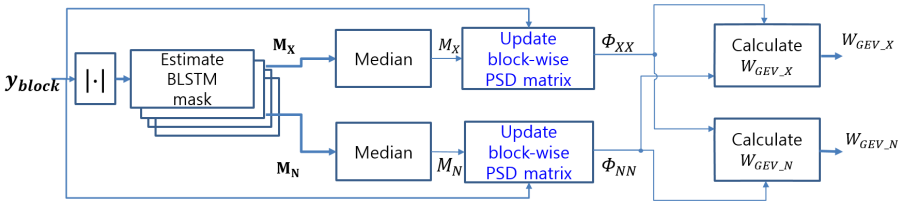

Fig. 3. (Color available online) Block diagram of beamforming vector estimator.

가중치 $\alpha_V^l(f)$는 블록 평균 마스크 값 $M_V^l(f)$를 이용해 아래 Eq. (11)과 같이 정의되며, 상수 r은 적합률로서 0에 가까우면 과거의 정보를 이용하지 않는 의미가 되며, 무한대가 되면 $\alpha_V^l(f)$이 0이 되어 현재 블록 정보가 반영되지 않는다.

$$\alpha_v^l(f) = \frac{M_v^l(f)}{M_v^l(f) + r}. \tag{11}$$

l번째 블록에서 추정된 $M_V^l(f)$는 Eq. (12)와 같이 BLSTM의 프레임 단위 마스크 추정 값 $M_V^l(t,f)$을 블록의 프레임 길이 L동안 합산한 후 평균 취한 값을 이용하여 얻어진다. 음성이 존재하지 않는 블록 구간에서는 음성 적응 가중치 $\alpha_X^l(f)$의 값이 작게 추정되어 음성 PSD 업데이트 시 적은 양을 적응한다.

$$M_v^l(f) = \sum_{t=1}^{L} M_v^l(t,f) / L. \tag{12}$$

3.1.2 PSD 업데이트를 위한 링버퍼 사용

Eq. (10)에서 구해진 PSD 행렬은 길이가 K인 링버퍼로 입력되며, 링 버퍼가 차면 계산된다. 최종적으로 Eq. (13)과 같이 계산되며 이를 이용해 빔포밍 벡터를 계산한다. 링 버퍼를 사용하는 이유는 현재 블록의 정보를 최대한 많이 반영하고, 과거 블록의 PSD 행렬 정보를 망각하기 위해서이다. 또한 블록 단위 업데이트 시 블록의 길이를 짧게 할수록 빔포밍 벡터가 계산되는 시점과 현재 입력 프레임 사이의 시간 지연(delay)이 블록 사이즈에 비례해서 줄어들지만, 빔포밍 벡터 계산을 위한 충분한 양의 프레임을 모을 수 없으면 PSD 행렬 계산 시 특정 주파수 빈의 샘플 수가 부족하면 부정확한 빔포밍 값이 나

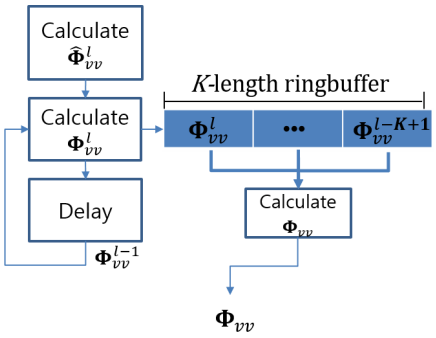

Fig. 4. (Color available online) Block diagram of block-wise PSD matrix update algorithm.

올 수 있다. 이러한 위험성을 줄이며, 빔포밍 벡터가 적용되는 시점에서 현재 입력 프레임에서 가까운 잡음 음성 정보를 최대한 반영하여 빔포밍하기 위해 링버퍼를 사용하였다. 제안하는 빔포밍 벡터 추정기 블록도 및 블록 단위 PSD 행렬 업데이트 알고리듬의 블록도는 Figs. 3과 4에서 확인할 수 있다.

$$\Phi_{vv}(f) = \sum_{i=0}^{K-1} \beta_i \Phi_{vv(i)}^{l-i}(f), \quad v \in \{X, N\}. \tag{13}$$

Fig. 3의 y_{block}은 Eq. (14)와 같이 정의되며, 블록 단위 업데이트 알고리듬에 사용되는 한 블록 안에 포함된 L개의 관측신호 프레임들을 의미한다. L개의 프레임들을 포함한 y_{block}을 입력으로 하여 음성 강화 빔포밍 벡터 W_{GEV_X}와 잡음 강화 빔포밍 벡터 W_{GEV_N}가 계산된다.

$$y_{\text{block}} = [y(t), ..., y(t-L+1)]. \tag{14}$$

3.2 스펙트럼 감산 빔포밍 시스템

빔포밍과 스펙트럼 감산을 결합한 선행 연구들이

시도되어 왔다.[18-21] 단일 채널 잡음 감소 기법인 스펙트럼 감산과 다채널 음성 강화 기법인 빔포밍은 상호 보완적인 효과를 볼 수 있다. 빔포밍 이전에 채널 별 스펙트럼 감산을 하게 되면 신호대잡음비 향상 효과를 거둘 수 있다. 또한 스펙트럼 감산에서 발생하는 신호 왜곡 현상인 음악적 잡음(musical noise)이 빔포밍으로 감소되는 효과가 있다.[20] 스펙트럼 감산을 빔포밍에 적용하는 방법은 Fig. 5와 같이 두 가지 방법이 연구되었다. 첫 번째는 Fig. 5(a)와 같이 빔포밍 후에 스펙트럼 감산을 적용하는 방법이다. 두 번째는 Fig. 5(b)와 같이 채널 별 스펙트럼 감산 후 빔포밍을 하는 방법이다. 기존 연구 결과로는 채널 별 스펙트럼 감산 후 빔포밍을 하는 방법이 잡음의 종류가 가우시안 및 슈퍼 가우시안 일 때, 더욱 좋은 잡음 제거 효과를 보였다.[20]

제안하는 알고리듬 또한 채널 별 스펙트럼 감산 후 빔포밍을 하는 방법을 사용한다. 기존 연구에서

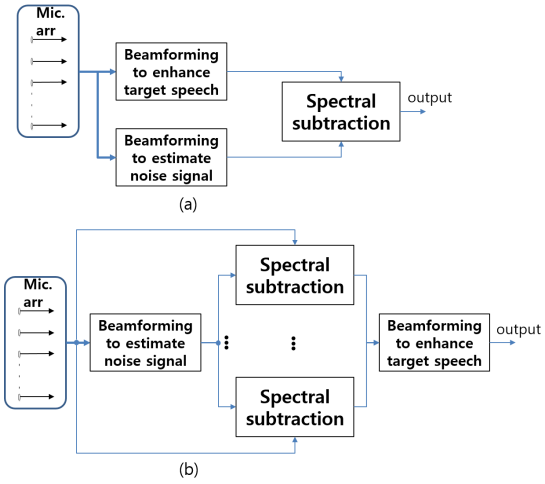

Fig. 5. (Color available online) Implementation methods of combined spectral subtraction and beamforming. (a) Spectral subtraction after beamforming, (b) beamforming after spectral subtraction.

는 음성 강화를 위해 Delay-and-Sum(DS) 빔포밍과 잡음 추정 기법으로는 원치 않는 방향의 간섭 신호를 제거하는 널조향(null-steering) 빔포밍 기법[22]을 활용하였다. 제안 알고리듬에서는 GEV의 빔포밍의 특성을 활용하여 GEV 빔포밍으로 음성 강화와 잡음 추정 빔포밍 벡터를 동시에 추정한다.

3.2.1 제안 스펙트럼 감산 온라인 빔포밍 시스템

본 연구에서 사용한 시스템은 GEV 빔포밍 시스템이 주파수 빈 별 신호대잡음비 최대화 알고리듬이기 때문에 Eq. (7)에서 목표 신호 PSD 행렬과 잡음 신호 PSD 행렬의 자리를 바꾸어 주면 잡음을 추정하는 빔포밍 벡터를 얻을 수 있다. Fig. 6은 스펙트럼 감산 온라인 빔포밍 시스템의 전체 블록도이다. 음성 강화 빔포밍 벡터 W_{GEV_X}와 잡음 강화 빔포밍 벡터 W_{GEV_N}가 동시에 추정되며, 추정된 W_{GEV_N}는 다채널 관측 신호 스펙트럼 $Y(t,f)$에 곱해져 잡음 추정에 사용된다. 각 채널에서 Eq. (15)와 같이 관측 신호 스펙트럼 크기 $|Y(t,f)|$에서 추정된 잡음 스펙트럼 크기 $|\hat{N}(t,f)|$를 감산하여 추정된 향상 신호 스펙트럼 크기를 $|\hat{S}(t,f)|$ 추정한다. 이때 각 채널에서 감산되는 추정 잡음 스펙트럼은 GEV 빔포머의 결과로 얻어진 동일한 추정 잡음 스펙트럼이다.

$$\left|\hat{S}(t,f)\right| = \left|Y(t,f)\right| - \left|N(t,f)\right| * \eta(t,f). \tag{15}$$

위 식의 η는 감산 가중치를 의미하며 관측 신호 스펙트럼 크기와 추정 잡음 스펙트럼 크기에 따라 Eq. (16)과 같이 정의하였다.

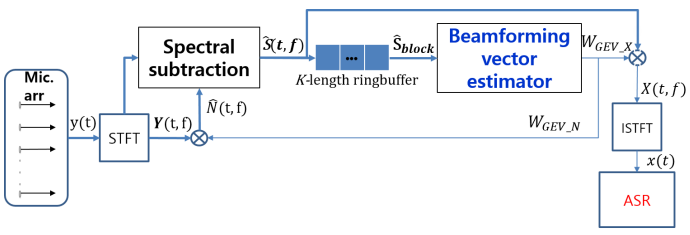

Fig. 6. (Color available online) Block diagram of spectral subtraction beamformer.

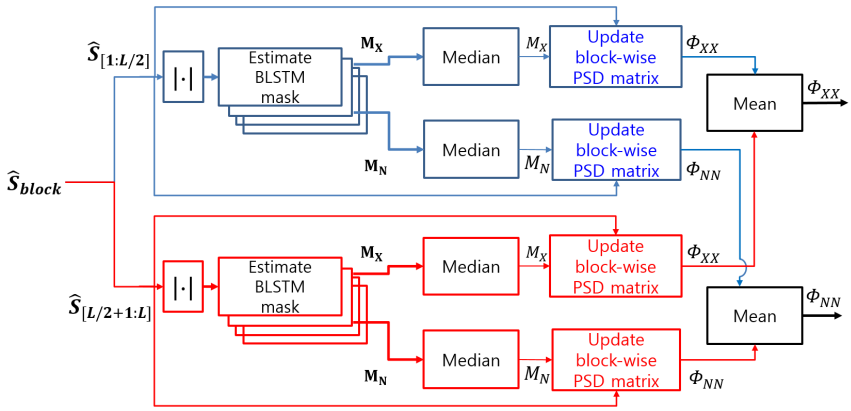

Fig. 7. (Color available online) Implementation of beamforming vector estimator by multiprocessing.

$$\eta(t,f)=\begin{cases}|Y(t,f)|/|\hat{N}(t,f)|*\lambda_N(f), & |Y(t,f)|<|\hat{N}(t,f)|\\ |Y(t,f)|-|\hat{N}(t,f)|)/|\hat{N}(t,f)|*\lambda_N(f), & \text{otherwise}\end{cases}. \quad (16)$$

$\lambda_N(f)$는 Eq. (17)과 같이 잡음 마스크 추정 값을 사용하여 정의되며 블록 단위 빔포밍 벡터 업데이트에서 사용된 L개의 프레임 중 시간적으로 현재에 가까운 $L/2$개의 프레임만을 사용하여 계산한다. 이렇게 하는 이유는 현재 시점의 프레임에 포함된 잡음의 주파수 빈 별 분포를 최대한 고려해서 감산하기 위해서이다. L개의 프레임을 사용할 시 적용 시점의 잡음 분포와 유사도가 떨어져서 성능 저하로 이어진다.

$$\lambda_N(f)=\sum_{t=L/2+1}^{L} M_N(t,f)/(L/2). \quad (17)$$

추정된 향상 신호 스펙트럼 복원을 위해서, 추정된 향상 신호 스펙트럼의 위상 정보 또한 필요하다. 짧은 구간의 위상 정보는 상대적으로 중요하지 않기 때문에,[20] 관측 신호 스펙트럼의 위상 정보를 고려하여 Eq. (18)과 같이 복원된다.

$$\hat{S}(t,f)=|\hat{S}(t,f)|e^{j\angle Y(t,f)}. \quad (18)$$

이렇게 스펙트럼 감산으로 추정된 향상 신호 스펙트럼 $\hat{S}(t,f)$가 길이가 L인 링버퍼로 들어가고 링버퍼가 차면 Eq. (19)와 같이 블록 단위의 향상된 신호 $\hat{S}_{\text{block}}$가 빔포밍 벡터 추정기의 입력으로 들어간다.

$$\hat{S}_{\text{block}}=[\hat{S}(t,f),...,\hat{S}(t-L+1,f)]. \quad (19)$$

최종적으로 Eq. (20)과 같이 빔포빙 벡터 추정기의 출력 W_{GEV_X}와 $\hat{S}(t,f)$의 곱으로 향상된 음성신호 스펙트럼 $X(t,f)$가 출력된다.

$$X(t,f)=\hat{S}(t,f)W_{GEV\_X}. \quad (20)$$

$X(t,f)$는 Inverse Short Time Fourier Transform(ISTFT)되고 시간 영역 신호 $x(t)$로 변환되어 음성인식기에 입력된다.

3.3 PSD 행렬의 빠른 수렴을 위한 블록 처리

블록단위 PSD 행렬 업데이트 시에, PSD 행렬의 빠른 수렴을 위해 $\hat{S}_{\text{block}}$를 절반으로 나누어 병렬처리한다. Fig. 7과 같이 나뉜 블록 배치 각각에 대해 PSD행렬 업데이트를 진행한다 (y_{block}으로 대체될 수 있음). 나누어진 배치 각각은 잡음 및 음성 PSD 행렬 계산에 사용되며, 각 배치로 계산된 PSD 행렬에 평균을 취하여 최종적인 Φ_{XX}와 Φ_{NN}을 추정한다. 후에는 기존과 같은 절차로 GEV 빔포밍 벡터를 계산한다.

이와 같이 블록을 나누어 처리하는 이유는 연속된 음성을 처리하는 온라인 빔포밍 테스크에서 PSD 행렬의 빠른 수렴이 성능 향상에 중요하기 때문이다. 블록을 나누어 각각의 PSD 행렬을 업데이트 후 평균 취하는 것이 전체 블록을 이용해 PSD 행렬을 한 번 업데이트하는 것보다 빠른 수렴으로 이어지는 효과를 기대하기 때문이다.

IV. 실험 결과

4.1 CSE 데이터베이스

실제 환경에서의 빔포밍 성능 평가를 위해, CHiME3의 개발 및 평가 시뮬레이션 세트와 MUSDB의 악기 및 보컬 음악을 혼합하여 CHiME3-MUSDB CSE 평가 세트를 제작하였다. CSE 평가 세트는 CHiME3 개발 시뮬레이션 발화 dt05 1,640개, 평가 시뮬레이션 발화 et05 1,320개를 목표 음성 신호로 사용하고, CHiME3의 버스 주행, 카페, 거리 교차로, 보행자 거리 잡음 및 MUSDB의 악기 와 보컬 음악 신호를 잡음 신호로 5종 잡음을 이용해 구성하였다.

5종 잡음 중 보컬 음악 잡음 세트는 총 12세트로 구성되며, 1 ~ 11번 세트는 250개의 무작위 시간 간격(3 s ~ 16 s 사이)을 두며 연쇄한 발화 단위 목표 음성 신호와, 목표 음성 신호 길이에 해당하는 MUSDB의 악기와 보컬 음악 신호를 연쇄하여 만들어진 잡음 신호를 혼합하여 6채널 잡음 음성 오디오 신호로 구성된다. 12번 세트는 이전 세트를 구성하고 남은 210개의 CHiME3 목표 음성 신호 발화를 이용해 같은 방식으로 만들어졌다.

나머지 버스 주행, 카페, 거리 교차로, 보행자 거리의 4종 배경 잡음은 CHiME3 챌린지에서 녹음된 연속 배경 잡음을 활용하였고 보컬 음악 잡음 세트와 같은 방식으로 만들어졌으며, 잡음별로 3세트로 구성되며 1 ~ 2번 세트는 250개 3번 세트는 240개의 발화 단위 목표 음성 신호와 그 길이에 해당하는 잡음을 혼합하여 만들어졌다. 총 24세트로 구성되며 모든 세트는 webrtcvad[23]를 사용해 목표 음성 신호에서 음성 존재 구간 정보를 얻어내고 이를 이용해 −10 dB, −5 dB, 0 dB, 5 dB, 10 dB의 신호대잡음비 별 평가 세트를 구성하였다.

추가적으로 한국어에 대한 성능 평가를 위해, 자체 제작된 버스 주행 잡음 환경 명령어(실제 도로 주행 버스 잡음 및 라디오 재생 환경에서 발성된 명령어)를 사용해 버스 주행 환경 CSE 평가 세트를 구성하였다. 버스 환경 한국어 평가 데이터는 4명의 화자(남자2, 여자2)로 구성되었고, 화자 당 100개의 명령어 발화를 사용하여 구성되었다. 총 2세트로 구성되었으며, 세트당 200발화를 포함한다.

CHiME3-MUSDB CSE 평가 세트와 같은 방식으로 발화 단위 목표 음성 신호와, 목표 음성 신호 길이에 해당하는 버스 주행 잡음을 사용하여 6채널 잡음 음성 오디오 신호를 생성하여, 신호대잡음비 별 평가 세트를 구성하였다.

4.2 마스크 추정을 위한 학습 및 개발 데이터베이스

BLSTM 마스크 추정 학습은 CHiME3의 학습, 개발 데이터베이스와 자체 제작된 차량 주행 잡음 환경 명령어(한국 가요, 엔진 잡음, 풍절음, 비상등 환경에서 발성된 명령어 녹음), 자체 제작된 버스 주행 잡음 환경 명령어(실제 도로 주행 버스 잡음 및 라디오 재생 환경에서 발성된 명령어 녹음)를 이용하여 학습되었다.

4.3 음성 강화 결과

빔포밍을 이용한 음성 강화의 결과는 CHiME3 챌린지[14]의 Kaldi 기반의 DNN 음향모델과 RNN 언어모델로 구성된 음성인식기를 이용한 단어오류율(Word Error Rate, WER) 평가를 진행하였다. CHiME3-MUSDB CSE 평가 세트 24개(음악혼합 음성 12세트/CHiME3 잡음 혼합 12세트)를 5개의 신호대잡음비 환경 총 120개의 연속 잡음 음성 신호 wav 파일에 대해 온라인 빔포밍 처리 후 얻어진 연속된 빔포밍 음성 스트림 wav 파일에서 음성 구간만을 발화 단위로 분절하여, 기존 CHiME3 dt05/et05 평가 세트로 재구성하여 평가를 진행하였다. 한국어에 대한 음성 강화 성능은 Google Web Speech API[24] 음성인식기를 이용하여 평가했다. 버스 주행 환경 CSE 평가 세트 2개를 5개의 신호대잡음비 환경별로 구성한 10개의 연속 잡음 음성 신호 wav 파일들을 온라인 빔포밍 처리 후 음성 구간만을 발화 단위로 분절한 wav 파일로

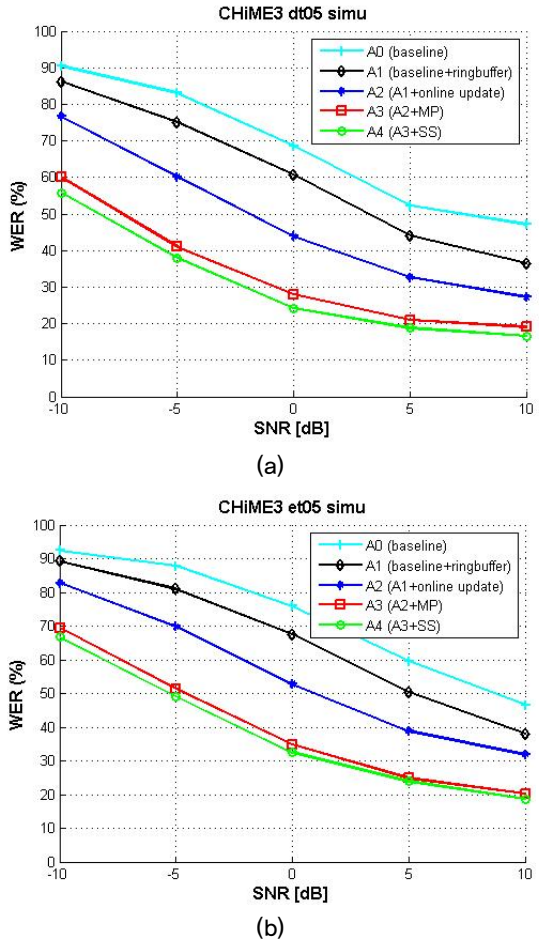

Fig. 8. (Color available online) WER performance of MUSDB mixing set. (a) CHiME3 dt05 set, (b) CHiME3 et05 set.

재구성 후에 평가를 진행하였다.

Figs. 8, 9에서 음악 잡음 음성 및 CHiME3 배경 잡음 음성에 대한 알고리듬별 평가 결과를 확인할 수 있다. 실험은 5 단계로 진행되었다. A0는 베이스라인으로 블록 단위로 PSD 행렬을 추정하여 빔포밍하는 방법이다. A1은 베이스 라인에 링버퍼를 추가한 방법이다. A2는 A1에 제안한 온라인 업데이트 알고리듬을 적용한 방법이다. A3는 A2에 빠른 수렴을 위한 블록 처리 후 업데이트를 진행한 방법이다. A4는 A3에 스펙트럼 감산을 적용한 방법이다. A0에서 A4로 제안 방법 및 알고리듬을 추가함에 따라 추가적인 성능 향상이 있었다.

A0에서 A1으로 성능 향상을 보면 링버퍼를 사용하는 것이 빔포밍시에 빔포밍 벡터 계산의 안정성을

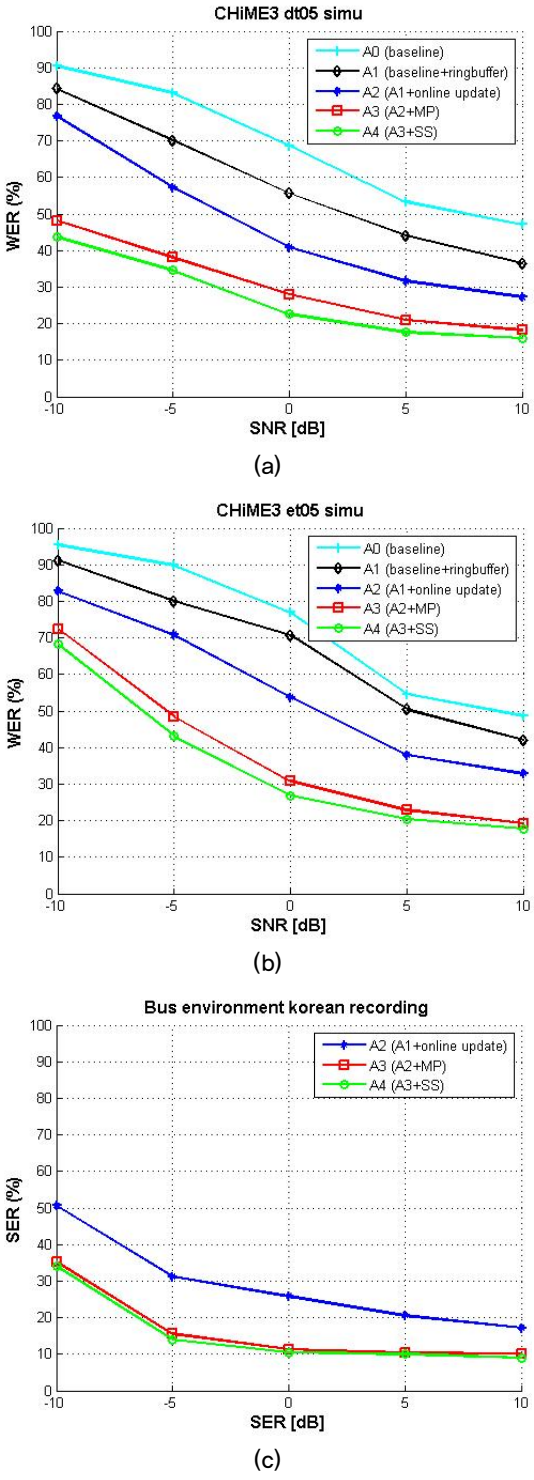

Fig. 9. (Color available online) WER performance of CHiME3 mixing set. (a) CHiME3 dt05 set, (b) CHiME3 et05 set.

확보하는 동시에 현재 입력 프레임과 빔포밍 계산시 사용되는 프레임들 사이의 지연을 늘리지 않아 성

Table 2. SNR (dB) measurements before and after spectral subtraction in CHiME3 dt05 set.

Input SNR (dB)	-10	-5	0	5	10
Before SS (Ch #1)	-9.8	-4.9	0.2	4.8	9.7
After SS (Ch #1)	-9.0	-4.0	1.0	4.5	9.3
After beamforming	-2.0	-1.8	-1.6	-1.5	-1.5

Table 3. PESQ measurements before and after spectral subtraction in CHiME3 dt05 set.

Input SNR (dB)	-10	-5	0	5	10
Before SS (Ch #1)	1.29	1.48	1.74	2.05	2.38
After SS (Ch #1)	1.37	1.62	1.85	2.06	2.37
After beamforming	1.81	2.02	2.25	2.37	2.48

능 향상으로 이어짐을 확인할 수 있었다. A2의 결과를 보면 제안한 온라인 업데이트 알고리듬이 BLSTM 마스크 추정 값을 이용해 음성이 존재하는 구간을 선별적으로 반영 후 업데이트하여 성능 향상으로 이어짐을 확인할 수 있었다. A3의 결과를 보면, PSD의 빠른 수렴을 위해 제안한 블록 처리 방법이 전체 배치를 업데이트하는데 비해 빠른 수렴으로 이어져 성능이 향상되었음을 알 수 있다. A4의 결과를 보면 BLSTM의 잡음 추정 값을 기반으로 스펙트럼 감산 방식 또한 빔포밍으로 들어가는 입력 잡음 음성의 신호대잡음비를 높이며 주며 음성 왜곡을 최소화하여 전반적인 성능 향상으로 이어졌음을 알 수 있다. 이를 증명하기 위해 A4의 빔포머 입력으로 들어가는 1번 채널 신호의 신호대잡음비 측정값이 Table 2에 제시되어 있다. A4의 빔포머 입력 신호는 스펙트럼 감산이 적용되어 1번 채널 입력 신호에 비해 신호대잡음비가 0 dB 이하 일 때는 높아지며, 5 dB 이상에서는 낮아지는 현상을 보인다. Table 3은 음성 품질의 평가 척도인 Perceptual Evaluation of Speech Quality(PESQ)의 측정값을 보여준다. 스펙트럼 감산 전후에서, 신호대잡음비가 0 이하일 때는 PESQ가 높아지며 5 이상에서는 유지되는 것을 확인할 수 있다. 평가는 CHiME3 dt05 세트의 분절된 발화 wav에 대해 이루어졌다.

et05 평가 데이터의 성능이 dt05 평가 데이터 보다 대체로 떨어지는 것은 dt05의 평균 발화 길이가 더 길기 때문이다. 발화 길이가 길수록 발화에 수렴된 빔포밍 벡터가 출력되고 발화가 짧을수록 발화에 수렴된 빔포밍 벡터의 안정성이 떨어지기 때문에 성능

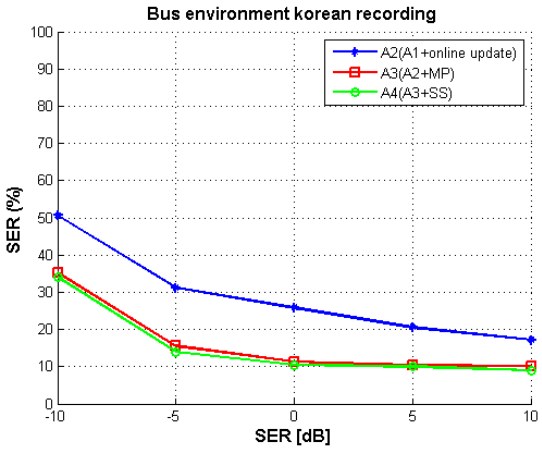

Fig. 10. (Color available online) SER of Korean recording set in bus environment.

Table 4. Samples of Korean commands used for evaluation. English translation is given in parentheses.

sample1	안치환의 사랑하게 되면 들려줘. (Play Ahn Chihwan's "If you love".)
sample2	마포아트센터 목적지로 설정. (Set destination to Mapo Art Center.)
sample3	구리점에 젠틀몬스터 팔아? (Does Guri Store sell Gentle Monsters?)
sample4	양재동 화훼공판장 안내해줘. (Guide me to Yangjae-dong Flower Market.)

열화로 이어지는 것을 확인할 수 있다.

Fig. 10은 한국어 버스 주행 잡음환경 평가 세트이며, 논문이 제안하는 알고리듬 A2, A3, A4에 대한 음절오류율(Syllable Error Rate, SER) 평가를 진행하였다. WER을 성능 지표로 사용할 경우, 한국어 평가 데이터의 특성상 복합명사를 포함하고 있어서 띄어쓰기 방법에 따라서 성능 변화가 크게 나타나기 때문에 SER를 이용한 평가를 진행하였다. 평가에 사용한 한국어 명령어 샘플들 중 일부를 Table 4에 제시하였다.

V. 결 론

신경망 빔포밍을 기반으로 블록 단위 온라인 빔포밍 알고리듬을 제안하였다. 실제 환경을 고려한 음성이 희박한 연속된 잡음 음성 빔포밍 테스크에서 성능 평가를 진행하였다. 실제 환경 테스트를 위해

CHiME3와 MUSDB 및 버스 환경 주행 잡음을 이용한 CSE 평가 세트를 제작하였으며, 테스크에 적합한 온라인 빔포밍 알고리듬 및 스펙트럼 감산 방법을 제안하여 효용성을 증명하였다.

제안한 온라인 빔포밍 알고리듬은 음성인식기 뿐만 아니라 잡음 제거 및 음성 강화가 필요한 모든 모듈의 전처리기로 활용될 수 있다. 그러나 낮은 신호대잡음비 환경에서는 상용화하기 힘든 수준의 WER을 보이며, 발화 길이가 짧아짐에 따라 성능이 열화되는 수렴 속도 문제 등 개선의 여지가 많이 남아있다. 향후 연구로서, 이러한 문제를 해결하기 위해 근본적인 지연 문제가 있는 BLSTM 기반 마스크 추정을 대체 및 보완하며 전반적인 성능 향상을 할 수 있는 딥러닝 모델 연구가 필요하다.

감사의 글

이 연구는 2018년도 산업통상자원부 및 산업기술평가관리원(KEIT) 연구비 지원에 의한 연구임(10080681).

References

1. S. Zhao, X. Xiao, Z. Zhang, T. N. T. Nguyen, X. Zhong, B. Ren, L. Wang, L. J. Douglas, E. Chng, and H. Li, “Robust speech recognition using beamforming with adaptive microphone gains and multichannel noise reduction,” Proc. IEEE Workshop on ASRU. 460-467 (2015).
2. Y. Tachioka, T. Narita, I. Miura, T. Uramoto, N. Monta, S. Uenohara, K. Furuya, S. Watanabe, and J. L. Roux, “Coupled Initialization of multi-channel non-negative matrix factorization based on spatial and spectral information,” Proc. 2017 INTERSPEECH, 2461-2465 (2017).
3. D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,” IEEE Trans. on Audio, Speech, and Lang. Process. **24**, 1626-1641 (2016).
4. T. V. d. Bogaert, S. Doclo, J. Wouters, and M. Moonen, “Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids,” J. Acoust. Soc. Am. **125**, 360-371 (2009).
5. E. A. Habets, J. Benesty, S. Gannot, and I. Cohen, “The MVDR beamformer for speech enhancement,” Proc. Speech Processing in Modern Communication, 225-254 (2010).
6. E. Warsitz and R. Haeb-Umbach, “Blind acoustic beamforming based on generalized eigenvalue decomposition,” IEEE Trans. on audio, speech, and lang. process. **15**, 1529-1539 (2007).
7. S. Gannot and I. Cohen, “Speech enhancement based on the general transfer function GSC and postfiltering,” IEEE Trans. on Speech and Audio Process. **12**, 561-571(2004).
8. J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, “BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge,” Proc. IEEE Workshop on ASRU. 444-451 (2015).
9. C. Deng, H. Song, Y. Zhang, Y. Sha, and X. Li, “DNN-based mask estimation integrating spectral and spatial features for robust beamforming,” Proc. ICASSP. 4647-4651 (2020).
10. Y. Liu, A. Ganguly, K. Kamath, and T. Kristjansson, “Neural network based time-frequency masking and steering vector estimation for two-channel MVDR beamforming,” Proc. ICASSP. 6717-6721 (2018).
11. N. Shankar, G. S. Bhat, and I. M. Panahi, “Real-time dual-channel speech enhancement by VAD assisted MVDR beamformer for hearing aid applications using smartphone,” Proc. 42nd Annual Int. Conf. of the IEEE EMBC. 952-955 (2020).
12. Y. Zhou, Y. Chen, Y. Ma, and H. Liu, “A real-time dual-microphone speech enhancement algorithm assisted by bone conduction sensor,” Sensors, **20**, 5050 (2020).
13. T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, “Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR,” IEEE Trans. on audio, speech, and lang. process. **25**, 780-793 (2017).
14. J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’speech separation and recognition challenge: Dataset, task and baselines,” Proc. 2015 IEEE Workshop on ASRU. 504-511 (2015).
15. Z. Rafii, A. Liutkus, F. R. Stöter, S. I. Mimilakis, and R. Bittner, MUSDB18 - a corpus for music separation (2017).
16. J. Heymann, L. Drude, and R. Haeb-Umbach, “Neural network based spectral mask estimation for acoustic beamforming,” Proc. IEEE ICASSP. 196-200 (2016).
17. E. Warsitz and R. Haeb-Umbach, “Blind acoustic beamforming based on generalized eigenvalue decomposition,” IEEE Trans. on audio, speech, and lang. process. **15**, 1529-1539 (2007).
18. J. S. Lim and A. V. Oppenheim, “Enhancement and

bandwidth compression of noisy speech," Proc. IEEE. 1586-1604 (1979).
19. D. Gala, A. Vasoya, and V. M. Misra, "Speech enhancement combining spectral subtraction and beamforming techniques for microphone array," Proc. the Int. Conf. and Workshop on Emerging Trends in Technology, 163-166 (2010).
20. Y. Takahashi, Y. Uemura, H. Saruwatari, K. Shikano, and K. Kondo, "Structure selection algorithm for less musical-noise generation in integration systems of beamforming and spectral subtraction," Proc. 2009 IEEE/SP 15th Workshop on Statistical Signal Processing, 701-704 (2009).
21. S. Karimian-Azari and T. H. Falk, "Modulation spectrum based beamforming for speech enhancement," Proc. 2017 IEEE WASPAA. 91-95 (2017).
22. H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," EURASIP J. Advances in Signal Processing, **2003**, 569270 (2003).
23. *Google WebRTC*, https://webrtc.org/, (Last viewed September 1, 2021).
24. *Google Web Speech API*, https://wicg.github.io/speech-api/, (Last viewed September 1, 2021).

저자 약력

▸ 윤 성 욱 (Sung-Wook Yoon)

2014년 2월 : 충북대학교 전자공학부 학사
2017년 2월 : 충북대학교 제어로봇공학전공 석사
2017년 3월 ~ 현재 : 충북대학교 제어로봇공학전공 박사 과정

▸ 권 오 욱 (Oh-Wook Kwon)

1986년 2월 : 서울대학교 전자공학과 학사
1988년 2월 : 한국과학기술원 전기및전자공학과 석사
1997년 2월 : 한국과학기술원 전기및전자공학과 박사
1988년 3월 ~ 2000년 4월 : 한국전자통신연구원 책임연구원
2000년 5월 ~ 2001년 3월 :한국과학기술원 연구교수
2001년 3월 ~ 2003년 8월 : UCSD 박사후 연구원
2003년 9월 ~ 현재 : 충북대학교 지능로봇공학과 교수, 컴퓨터정보통신연구소 교수