한국음향학회지 제41권 제1호 pp. 38~44 (2022)
The Journal of the Acoustical Society of Korea Vol.41, No.1 (2022)
https://doi.org/10.7776/ASK.2022.41.1.038

pISSN : 1225-4428
eISSN : 2287-3775

효과적인 복소 스펙트럼 기반 음성 향상을 위한 시간과 주파수 영역 손실함수 조합에 관한 연구

A study on loss combination in time and frequency for effective speech enhancement based on complex-valued spectrum

정재희,[1] 김우일[1†]

(Jaehee Jung[1] and Wooil Kim[1†])

[1]인천대학교 컴퓨터공학부

(Received November 26, 2021; revised December 28, 2021; accepted January 10, 2022)

초 록: 잡음에 오염된 음성의 명료도와 음질을 향상시키고자 음성 향상을 수행한다. 본 연구에서는 복소값 스펙트럼을 이용한 마스크기반 음성 향상에서 시간 영역 손실함수와 주파수 영역 손실함수에 따른 학습 결과를 비교하였다. 시간 영역의 음성 파형과 주파수 영역의 스펙트럼의 세부정보를 고려해 두 영역의 장점을 활용할 수 있도록 손실함수 조합에 관해 연구를 진행하였다. 시간 영역 손실함수는 Scale Invariant-Source to Noise Ratio(SI-SNR)을 이용해 계산하고, 주파수 영역 손실함수는 복소값 스펙트럼과 크기 스펙트럼을 Mean Squared Error(MSE)로 계산하여 사용하였고, sin 함수를 이용해 위상에 대한 손실함수를 계산하였다. 손실함수 조합은 시간 영역 손실함수인 SI-SNR과 각 주파수 영역 손실함수를 조합하였다. 또한 크기 값과 위상 값을 모두 고려할 수 있도록 SI-SNR과 크기 스펙트럼, 위상에 관련된 손실함수들도 조합하여 실험을 진행하였다. 음성 향상 결과는 Source-to-Distortion Ratio(SDR), Perceptual Evaluation of Speech Quality(PESQ), Short-Time Objective Intelligibility(STOI)를 이용해 성능 비교 평가를 진행하였다. 음성 향상 결과를 확인해보기 위해 스펙트럼 상에서 비교를 진행하였다. TIMIT 데이터베이스를 이용한 실험 결과, 시간 영역 또는 주파수 영역 손실함수보다 SI-SNR과 크기 스펙트럼을 조합한 손실함수를 사용하여 음성 향상을 학습했을 때 가장 높은 성능을 보였다.

핵심용어: 복소값 스펙트럼, 음성 향상, 시간 영역, 주파수 영역, 손실함수 조합

ABSTRACT: Speech enhancement is performed to improve intelligibility and quality of the noise-corrupted speech. In this paper, speech enhancement performance was compared using different loss functions in time and frequency domains. This study proposes a combination of loss functions to utilize advantage of each domain by considering both the details of spectrum and the speech waveform. In our study, Scale Invariant-Source to Noise Ratio (SI-SNR) is used for the time domain loss function, and Mean Squared Error (MSE) is used for the frequency domain, which is calculated over the complex-valued spectrum and magnitude spectrum. The phase loss is obtained using the sin function. Speech enhancement result is evaluated using Source-to-Distortion Ratio (SDR), Perceptual Evaluation of Speech Quality (PESQ), and Short-Time Objective Intelligibility (STOI). In order to confirm the result of speech enhancement, resulting spectrograms are also compared. The experimental results over the TIMIT database show the highest performance when using combination of SI-SNR and magnitude loss functions.

Keywords: Complex-valued spectrum, Speech enhancement, Time domain, Frequency domain, Loss combination

PACS numbers: 43.72.Bs, 43.72.Ne

†**Corresponding author:** Wooil Kim (wikim@inu.ac.kr)
Department of Computer Science and Engineering, Incheon National University, 119 Academy-ro, Yeonsu-gu, Incheon 22012, Republic of Korea
(Tel: 82-32-835-8459, Fax: 82-32-835-0780)

"이 논문은 2021년도 추계학술대회에서 발표하였던 논문임."

I. 서 론

음성 향상은 잡음이 섞여 있는 음성에서 목표 음성의 명료도와 지각적인 품질을 올리는 것을 목표로 수행된다. 과거에는 음성 향상을 위해 통계적 기법[1,2] 기반의 연구가 주로 수행되어 왔으나, 최근에는 심층 신경망을 이용한 음성 향상 기법이 많이 연구되고 있다.[3,4] 심층 신경망 기반 음성 향상 기법 중 대표적인 것으로 목표 음성의 마스크를 추정하는 마스크 기반 방법이 있다. 마스크기반 음성 향상은 심층 신경망을 통해 마스크를 추정하고, 이를 입력에 사용된 음성 파형이나 스펙트럼에 곱해 향상된 음성 파형이나 스펙트럼을 얻는다. 이때 입력에 사용되는 음성 파형이나 스펙트럼의 종류에 따라 다양한 마스크기반 음성 향상으로 나뉠 수 있다.[5] 지금까지도 많이 사용되는 방법은 크기 스펙트럼을 이용한 방법으로 잡음 오염 음성의 크기 스펙트럼을 입력으로 사용한다.[6] 그 후 심층 신경망을 통해 추정된 마스크와 입력 스펙트럼을 곱해 향상된 음성의 크기 스펙트럼을 얻는다. 향상된 음성의 크기 스펙트럼은 잡음 오염 음성의 위상을 이용하여 향상된 음성 파형을 얻을 수 있다. 하지만 이 방법에서 잡음 오염 음성 위상을 그대로 사용하기 때문에 음성 향상에 한계가 있다.[7,8] 이에 따라 최근에는 스펙트럼의 크기와 위상을 모두 향상 시킬 수 있도록 Short Time Fourier Transform(STFT) 변환 후 복소값을 가지는 스펙트럼을 그대로 이용해 음성 향상을 수행하는 방법이 많이 연구되고 있다.[9,10]

본 연구에서는 복소값 스펙트럼을 이용한 음성 향상의 효과적인 학습을 위해 많이 사용되는 시간 영역 손실함수와 주파수 영역 손실함수에 따른 결과를 비교하였다. 또한 시간 영역의 음성 파형과 주파수 영역의 스펙트럼의 세부 정보를 모두 고려해 두 영역의 장점을 모두 학습에 이용할 수 있도록 각 영역의 손실함수를 조합하는 방안에 대해 연구를 진행하였다. 복소값 스펙트럼을 이용한 음성 향상을 위해 많이 사용되고 있는 모델 중 하나인 Deep Complex U-Net (DCUNET)에 complex attention이 적용된 모델을 이용하였다.[10,11] 사용한 시간 영역 손실함수는 Scale Invariant-Source to Noise Ratio(SI-SNR) 손실함수[12]를 사용하였고 주파수 영역 손실함수는 Mean Squared Error(MSE) 손실함수와 sin함수를 이용해 계산하였다.[13]

다음 II장에서는 음성 향상 기법에 대한 전체적인 과정과 사용한 모델에 대해 설명하고, III장에서는 사용한 손실함수와 손실함수 조합에 관해 설명한다. IV장에서는 실험과 그 결과에 대해 논의하고 V장에서 결론을 맺는다.

II. 마스크기반 음성 향상

2.1 음성 향상 특징

음성 향상에 사용되는 잡음 오염 음성은 다음과 같이 시간 영역 파형으로 생성한다.

$$y = s + n, \tag{1}$$

여기서 y는 생성된 잡음 오염 음성의 파형을 나타내고, s와 n은 각각 깨끗한 음성과 잡음을 나타낸다.

이렇게 생성된 잡음 오염 음성은 STFT를 통해 복소값 스펙트럼으로 변환해 음성 향상에 사용된다.

$$Y = Y_r + jY_i. \tag{2}$$

위의 식에서 Y는 잡음 오염 음성의 복소값 스펙트럼으로 실수부를 나타내는 r과 허수부를 나타내는 i로 구성된 행렬 형태를 갖는다. 또한 깨끗한 음성 스펙트럼 S 도 Eq. (2)와 같이 표현된다.

2.2 복소값 스펙트럼을 이용한 음성 향상

본 연구에서 수행한 이용한 음성 향상 과정은 Fig. 1과 같다. STFT를 통해 얻어진 복소값 스펙트럼은 심층 신경망 모델에 입력으로 사용되고, 모델은 마스크를 추정한다. 추정된 마스크는 입력으로 사용되었던 스펙트럼과 곱해져 향상된 음성의 스펙트럼을

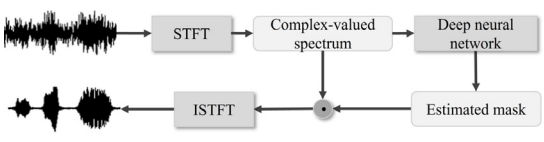


Fig. 1. Process of speech enhancement system using complex-valued spectrum.

얻을 수 있다. 시간 영역의 향상된 음성의 파형은 향상된 음성의 스펙트럼에 Inverse STFT(ISTFT)를 취해서 얻을 수 있다.

본 연구에서 추정된 마스크와 입력 스펙트럼의 곱은 다음과 같이 크기 스펙트럼과 위상 스펙트럼으로 변환하여 계산하였다.

$$\hat{S} = |Y| \cdot |M| \cdot e^{j(\theta_Y + \theta_M)}. \quad (3)$$

위 식에서 M은 추정된 마스크를 나타내며, $\hat{S}$은 향상된 음성의 스펙트럼을 나타낸다. 또한 $|\cdot|$는 크기 스펙트럼을, θ는 위상 스펙트럼을 나타낸다. 여기서 M과 $\hat{S}$ 모두 Eq. (2)처럼 실수부와 허수부로 구성된 복소값 행렬형태이다.

2.3 음성 향상 모델

본 연구에서는 음성 향상을 위한 모델로 많이 사용되는 Deep Complex U-Net(DCUNET) 모델에 Complex Attention(CA)모듈을 적용하여 사용하였다.[10,11] 사용한 모델의 구조는 Fig. 2와 같다.

모델은 Encoder 8계층, Decoder 8계층으로 구성되어 있고 Encoder와 Decoder의 각 층은 CA가 적용된 skip-connection으로 연결되어 있다. Encoder는 복소 콘볼루션, 배치 정규화, 활성화 함수로 구성되어 있고, Decoder는 Encoder와 유사한 구조이지만 복소 콘볼루션 대신 복소 전치 콘볼루션이 수행된다.

본 연구에서 복소 콘볼루션 연산 및 배치 정규화, 활성화 함수는 다음 식과 같이 적용된다.[10,14]

$$W = W_r + j W_i. \quad (4)$$

$$X = X_r + j X_i. \quad (5)$$

$$F_{conv} = (X_r * W_r - X_i * W_i) + j(X_r * W_i + X_i * W_r). \quad (6)$$

$$output = F_r(X_r) + j F_i(X_i). \quad (7)$$

위 식에서 W_r과 W_i는 서로 다른 콘볼루션 필터이다. 복소값 스펙트럼에 콘볼루션 연산을 적용하기 위해 복소 콘볼루션 필터 W는 Eq. (4)와 같이 정의된다. 모델의 중간 복소값 스펙트럼을 Eq. (5)에서 X라고 가정했을 때 복소 콘볼루션은 Eq. (6)과 같이 계산된다. 배치 정규화와 활성화 함수는 복소 콘볼루션과 다르게 Eq. (7)과 같이 실수부분, 허수부분 행렬에 각각 함수를 별도로 적용하는 방법으로 연산을 수행한다. 위의 Eq. (7)에서 F 함수는 배치 정규화 연산 또는 활성화함수가 될 수 있다.

모델에 적용된 CA는 Fig. 3과 같이 계산된다. 입력으로는 Encoder와 Decoder의 결과 값을 사용한다. 모듈의 연산 결과로 나온 값은 입력으로 사용했던 De-

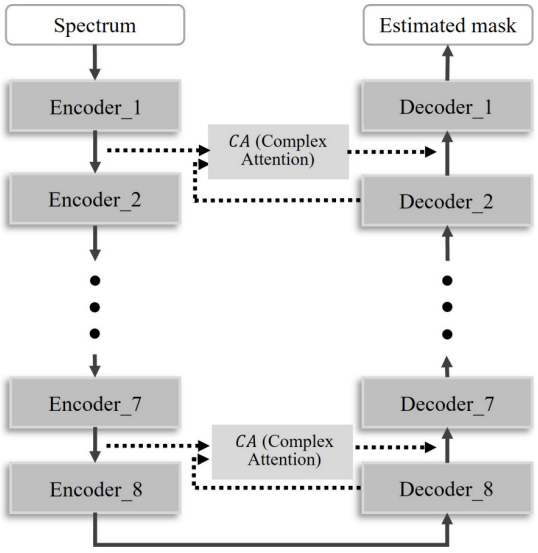


Fig. 2. Structure of DCUNET with complex attention.

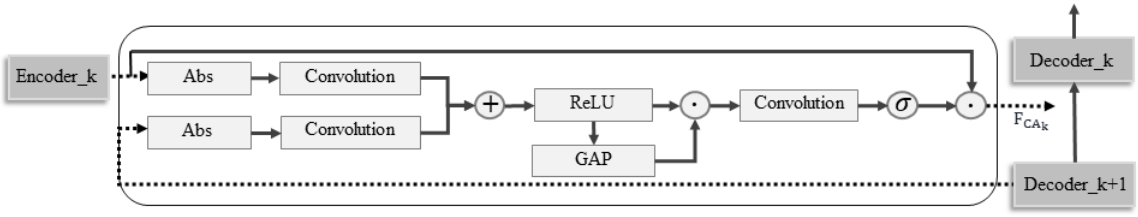


Fig. 3. Structure of complex attention module in DCUNET.

coder의 결과 값과 연결되어 다음 층의 Decoder의 입력으로 사용된다. Fig. 3에서 Encoder_k, Decoder_k 와 Decoder_k+1에 사용되는 k는 각 층을 나타내며, 1부터 7까지의 값이 될 수 있다. Fig. 3의 σ은 sigmoid 활성화 함수를 나타낸다. CA모듈에서 사용된 Global Average Pooling(GAP)연산이나 Absolute(Abs)연산은 모두 Eq. (7)과 같은 방식으로 실수부분, 허수부분 각각 연산이 적용된다.

III. 시간 영역, 주파수 영역 손실함수

음성 향상을 위한 손실함수는 복소값 스펙트럼을 이용한 음성 향상 연구에서 많이 사용된 손실함수를 이용하여 연구를 진행하였다.

3.1 시간 영역 손실함수

시간 영역 손실함수는 향상된 스펙트럼에 ISTFT를 취해 얻은 음성 파형을 이용해 계산한다. 본 논문에서는 시간 영역 손실함수로 SI-SNR 손실함수를 사용하였다. SI-SNR 손실함수인 L_{si-snr}은 크기가 조정된 깨끗한 음성과 그 외 잡음의 비율로 계산된다.

$$s_{target} = \frac{(<\hat{s}, s> \cdot s)}{\|s\|_2^2}. \tag{8}$$

$$e_{noise} = \hat{s} - s_{target}. \tag{9}$$

$$L_{si-snr} = 10\log_{10}(\frac{\|s_{target}\|_2^2}{\|e_{noise}\|_2^2}). \tag{10}$$

위 식에서 $<\cdot,\cdot>$ 은 두 벡터에 대한 내적 곱 연산을 나타내고, $\|\cdot\|_2$는 L2 정규화 연산을 나타낸다. $\hat{s}$은 향상된 음성 파형이고, s는 깨끗한 음성 파형이다.

3.2 주파수 영역 손실함수

주파수 영역 손실함수는 복소값 스펙트럼과 크기, 위상에 대해 MSE와 sin함수를 이용해 계산한다. 첫 번째는 복소값 스펙트럼을 이용해 깨끗한 음성의 스펙트럼과의 차이를 계산한다. 스펙트럼이 복소값을 가지기 때문에 깨끗한 음성의 스펙트럼과 향상 음성의 스펙트럼의 실수부, 허수부 각각에 대해 계산한다. 두 번째는 크기 값을 이용해 깨끗한 음성의 크기 값과의 차이를 계산한다. 세 번째로 위상은 각도와 관련된 값으로 직접적으로 MSE로 차이를 계산하지 않고 sin함수를 추가로 이용해서 손실함수를 계산할 수 있다.

먼저 MSE는 임의의 행렬 A와 B에 대해서 다음과 같이 두 행렬의 차의 제곱으로 계산되어진다.

$$MSE(A,B) = \frac{1}{TF}\sum_{t=0}^{T-1}\sum_{f=0}^{F-1}(A_{t,f} - B_{t,f})^2. \tag{11}$$

위 식에서 t와 f는 각각 시간 축 프레임 인덱스와 주파수 인덱스를 나타내고, T와 F는 각각 시간 축 프레임의 총 개수와 주파수 간격의 총 개수이다.

MSE를 이용해서 계산하는 복소값 스펙트럼과 크기 스펙트럼에 대한 손실함수는 다음과 같다.

$$L_{spectrum} = MSE(\hat{S_r}, S_r) + MSE(\hat{S_i}, S_i). \tag{12}$$

$$L_{magnitude} = MSE(|\hat{S}|, |S|). \tag{13}$$

위 식에서 $\hat{S}$과 S는 향상된 음성의 스펙트럼과 깨끗한 음성의 스펙트럼으로 r과 i는 각 스펙트럼의 실수부와 허수부이다. Eq. (13)에서 $|\cdot|$는 크기 스펙트럼을 나타낸다. $L_{spectrum}$이 복소값 스펙트럼을 이용한 손실함수이고 $L_{magnitude}$가 크기 스펙트럼을 이용한 손실함수이다.

sin함수를 이용하여 스펙트럼의 위상인 θ에 대한 손실함수는 다음과 같이 계산한다.[13]

$$L_{phase} = \frac{1}{TF}\sum_{t=0}^{T-1}\sum_{f=0}^{F-1}(\sin(\frac{\theta_{\hat{S_{t,f}}} - \theta_{S_{t,f}}}{2}))^2. \tag{14}$$

3.3 손실함수 조합

본 논문에서는 복소값 스펙트럼을 이용한 음성 향상에서 시간 영역 파형 정보와 주파수 영역 스펙트럼의 세부 정보를 같이 고려해 학습할 수 있도록 두

영역의 손실함수를 조합하였다. 또한 본 연구에서 실험을 통해 각 영역 손실함수의 학습 결과를 관찰하였을 때, Fig. 4(c)에서 볼 수 있는 시간 영역 손실함수를 사용한 향상은 잡음을 제거하는 것보다 음성 손상을 최소화하는 것에 중점을 두고 향상을 수행하는 학습결과를 보였다. 반면에 Fig. 4(d)에서 볼 수 있는 주파수 영역 손실함수를 사용한 향상 결과는 잡음을 보다 확실히 제거하지만 음성도 많이 제거되는

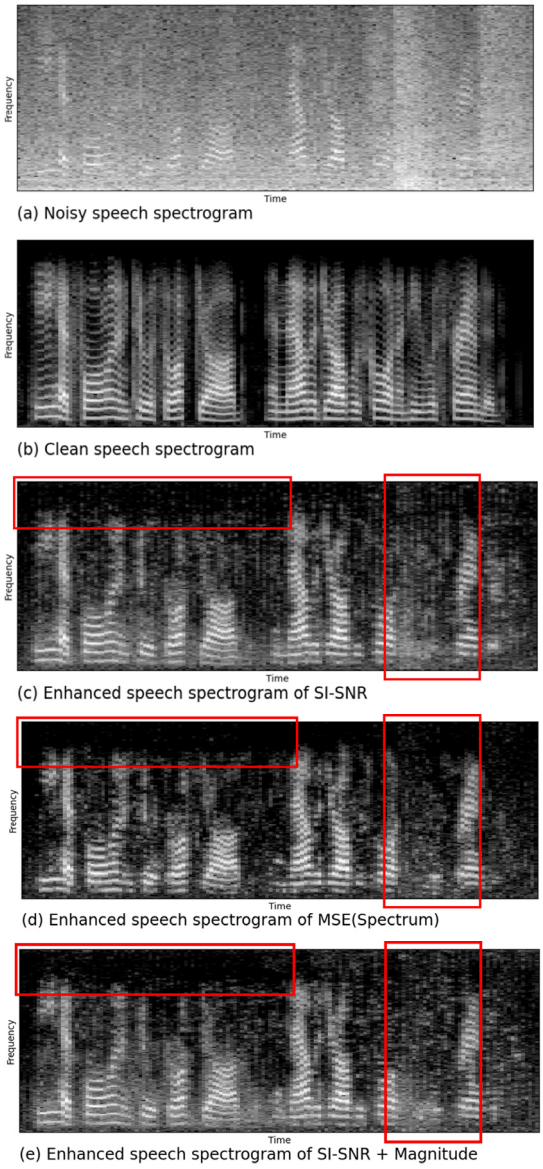


Fig. 4. (Color available online) A spectrogram is one of the test samples. The sample is a 'factory' noise type with SNR −5dB. The spectrogram is in a log-scale.

경향을 보인다. 따라서 이 두 손실함수를 같이 조합함으로써 음성 손상을 최소로 유지하면서 잡음을 보다 확실히 제거할 수 있음을 기대할 수 있다.

손실함수 조합은 시간 영역 손실함수인 SI-SNR 손실함수와 주파수 영역 손실함수들을 각각 조합하고 추가로 극좌표계의 크기와 위상을 한 번에 고려할 수 있도록 두 손실함수를 같이 조합하였다. 조합한 손실함수는 총 4가지로 다음과 같다.

$$L_{SI-SNR+Spectrum} = L_{si-snr} + L_{spectrum}. \quad (15)$$

$$L_{SI-SNR+Magnitude} = L_{si-snr} + L_{magnitude}. \quad (16)$$

$$L_{SI-SNR+Phase} = L_{si-snr} + L_{phase}. \quad (17)$$

$$L_{SI-SNR+Magnitude+Phase} = L_{si-snr} + L_{magnitude} + L_{phase}. \quad (18)$$

IV. 실험 및 결과

본 논문에서 사용한 깨끗한 음성은 TIMIT 데이터베이스[15]를 사용하였고 잡음에는 NOISEX92와 대중가요를 사용하였다.[16] 잡음 종류는 총 4가지로 'Factory', 'Car', 'Babble', 'Music'이고 깨끗한 음성과 SNR 5 dB, 0 dB, -5 dB의 조건으로 합하여 잡음 오염 음성을 생성하였다. 혼합된 잡음 오염 음성은 훈련에 55,440발화, 검증에는 1,200발화, 테스트에는 2,304발화를 사용하였다. 음성은 모두 8 KHz이며, 윈도우 크기는 64 ms로 50 % 씩 겹치게 이동하였다. 푸리에 변환은 512개로 설정하여 주파수 256차원과 에너지 값을 포함해 총 257차원으로 설정하였다.

훈련을 위한 최적화 알고리즘은 'Adam'을 사용하였고 학습률은 0.001로 설정하였다. 모델에 사용된 활성화함수는 디코더의 마지막 층을 제외하고 모두 Leaky-ReLU 함수를 사용하였고, 마지막 층은 Tanh 함수를 적용하였다. 실험은 같은 환경에서 동일한 횟수의 학습을 진행하여, 시간 영역 손실함수와 주파수 영역 손실함수를 적용한 모델과 두 영역의 손실함수를 조합하여 적용한 모델들에 대해 비교를 진행하였다.

모델에 대한 성능 평가는 총 3가지 지표를 이용해 비교하였다. 첫 번째는 깨끗한 음성과 원하지 않은 왜곡의 비율을 나타내는 SDR로 dB단위로 표현된다.[17] 두 번째는 PESQ로 음성의 지각적인 품질을 평가하는데 가장 많이 사용되며 ITU-T에서 표준화하였다.[18] PESQ 지표는 보통 대부분 1.0에서 4.5사이의 값을 가진다. 세 번째는 음성의 명료도를 객관적으로 평가할 수 있는 STOI을 이용하였다.[19] STOI는 0에서 1사이의 값을 가지기 때문에 백분율로 단위를 조정해서 비교를 진행하였다. 또한 추가로 객관적인 지표뿐 아니라 실제 결과를 직접 비교할 수 있도록 Fig. 4에 시간 영역, 주파수 영역 손실함수와 이를 조합한 손실함수로 훈련된 모델의 결과 음성을 스펙트로그램으로 나타내 분석하였다.

객관적인 지표를 이용한 성능 평가 결과를 Table 1에 나타내었다. 표에서 No processing은 음성 향상을 수행하지 않은 잡음 오염 음성에 관한 평가 결과이다. 먼저 시간 영역 손실함수인 SI-SNR과 주파수 영역 손실함수인 Spectrum을 비교하였을 때, 시간 영역 손실함수를 사용한 경우가 더 좋은 음성 향상 결과를 나타내었다. 이는 Fig. 4에서 차이를 확인할 수 있다. Fig. 4에 사용된 음성은 테스트 음성 중 하나로 (a)는 잡음 오염 음성의 스펙트로그램, (b)는 깨끗한 음성의 스펙트로그램을 나타낸다. (c)는 시간 영역 손실함수로 훈련된 모델의 음성 향상 결과를 나타내고 (d)는 주파수 영역 손실함수로 훈련된 모델의 음성 향상 결과를 나타낸다. Fig. 4(c)와 (d) 스펙트럼 중 표시한 박스 부분을 비교해봤을 때, (c)는 잡음을 덜 제거시키지만, 음성이 제거되는 부분도 적도록 향상을 수행한다. 반면에, (d)에서는 음성이 손상되더라도 잡음을 보다 확실히 제거한 향상 결과를 보여준다. 본 연구와 같은 극심한 잡음이 있는 환경에서는 Fig. 4(d) 경우에서 오히려 잡음 제거만을 중점으로 진행하면서 중요한 음성 부분들이 사라져 (c)와 성능 차이가 난 것으로 볼 수 있다.

Table 1. Result of speech enhancement trained with various loss function. The results are averaged for all test data set.

Loss function	SDR	PESQ	STOI
No processing	0.56	1.629	65.9
SI-SNR	15.46	2.907	88.7
Spectrum	9.86	2.725	82.3
SI-SNR+Spectrum	14.63	2.869	87.8
SI-SNR+Magnitude	**15.50**	**2.912**	**88.7**
SI-SNR+Phase	15.27	2.885	87.9
SI-SNR+Magnitude+Phase	15.09	2.894	88.0

다음으로 시간 영역의 파형과 스펙트럼의 세부정보를 같이 고려한 두 영역의 손실함수 조합 결과를 비교하였다. 그 결과, 복소값 스펙트럼의 모든 정보보다는 크기 정보만을 고려한 Magnitude 손실함수와 SI-SNR 손실함수와 조합하였을 때, 가장 높은 성능을 보였다. 또한 이 손실함수 조합은 주파수 영역 손실함수만을 사용했을 때보다 SDR 5.64 dB, PESQ 0.187, STOI 6.4 %만큼 성능이 향상되었고, 시간 영역 손실함수만 사용했을 때보다도 SDR 0.04 dB, PESQ 0.005만큼 성능이 향상되었다. 본 연구에서는 손실함수에 대한 객관적 평가를 위해 모든 실험에서 동일한 횟수로 학습을 진행하였다. 일반적으로 손실함수 조합은 단일 손실함수보다 더 많은 학습을 필요로 하기 때문에 더 많은 학습을 진행한다면 제안한 손실조합에서 더 큰 성능 향상을 기대할 수 있다. 해당 손실함수 조합을 이용한 음성 향상 결과는 Fig. 4(e)에서 확인할 수 있다. Fig. 4에서 박스 안 스펙트럼 결과를 보면 (e)가 (c)에 비해 잡음이 더 확실히 제거되고 (d)보다는 음성의 손상을 줄인 결과를 확인할 수 있다.

V. 결 론

본 논문에서는 복소값 스펙트럼을 사용하는 음성 향상의 효과적인 학습을 위해 손실함수 조합에 관해 연구를 진행했다. 시간 영역 손실함수를 이용한 음성 향상은 음성의 손상을 최대한 줄이는 선에서 잡음을 제거한다. 반면에 주파수 영역 손실함수는 음성에 왜곡이 발생되더라도 확실히 잡음을 제거하도록 학습된다. 따라서 이 두 영역의 손실함수를 여러 조합으로 실험을 진행하였다. 그 결과, SI-SNR 손실함수와 스펙트럼 크기 값을 이용한 손실함수를 조합하였을 때, 각 영역의 손실함수를 사용했을 때보다 SDR, PESQ, STOI 모두 성능이 향상되었음을 확인하였다.

감사의 글

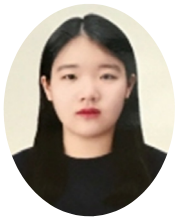
본 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2019R1F1A106299513).

References

1. J. Lim and A. Oppenheim, “All-pole modeling of degraded speech,” IEEE Trans. on Acoustics, Speech, and Signal Process. **26**, 197-210 (1978).
2. R. Martin, “Spectral subtraction based on minimum statistics,” power 6.8 (1994).
3. Y. H. Tu, J. Du, and C. H. Lee, “2d-to-2d mask estimation for speech enhancement based on fully convolutional neural network,” Proc. IEEE ICASSP. 6664-6668 (2020).
4. Y. Xu, J. Du, and C. H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Trans. on Audio, Speech, and Lang. Process. **23**, 7-19 (2014).
5. D. L. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” IEEE/ACM Trans. on Audio, Speech, and Lang. Process. **26**, 1702-1726 (2018).
6. Z. Xu, S. Elshamy, and T. Fingscheidt, “Using separate losses for speech and noise in mask-based speech enhancement,” Proc. IEEE ICASSP. 7519-7523 (2020).
7. K. Paliwal, K. Wójcicki, and B. Shannon, “The importance of phase in speech enhancement,” speech communication, **53**, 465-494 (2011).
8. Y. Wang and D. L. Wang, “A deep neural network for time-domain signal reconstruction,” Proc. IEEE ICASSP. 4390-4394 (2015).
9. Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, “DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement,” arXiv preprint arXiv:2008.00264 (2020).
10. H. S. Choi, J. H. Kim, J. Huh, A. Kim, J. W. Ha, and K. Lee, “Phase-aware speech enhancement with deep complex u-net,” Proc. ICLR. 1-20 (2019).
11. O. Oktay, J. Schlemper, L. Le. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, “Attention u-net: Learning where to look for the pancreas,” arXiv preprint arXiv:1804.03999 (2018).
12. Y. Luo and N. Mesgarani, “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation,” IEEE/ACM Trans. on Audio, Speech, and Lang. Process. **27**, 1256-1266 (2019).
13. J. Zhang, M. D. Plumbley, and W. Wang, “Weighted magnitude-phase loss for speech dereverberation,” Proc. IEEE ICASSP. 5794-5798 (2021).
14. C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, “Deep complex networks,” arXiv preprint arXiv:1705.09792 (2017).
15. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1,” NASA STI/Recon Tech. Rep., n 93: 27403, 1993.
16. A. Varga, “The NOISEX-92 study on the effect of additive noise on automatic speech recognition,” ical Report, DRA Speech Research Unit, CiNii (1992).
17. E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” IEEE Trans. on Audio, Speech, and Lang. Process. **14**, 1462-1469 (2006).
18. A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” Proc. IEEE ICASSP. 749-752 (2001).
19. C. H. Taal, R. C. Hendriks, and R. Heusdens, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” Proc. IEEE ICASSP. 4214-4217 (2010).

저자 약력

▸ 정 재 희 (Jaehee Jung)



2021년 2월 : 인천대학교 컴퓨터공학부 공학사
2021년 3월 ~ 현재 : 인천대학교 컴퓨터 공학과 석사과정

▸ 김 우 일 (Wooil Kim)

1996년 2월, 1998년 8월, 2003년 8월 : 고려대학교 전자공학과 학/석/박사
2012년 8월 ~ 현재 : 인천대학교 컴퓨터공학부 조교수, 부교수, 교수