한국음향학회지 제41권 제1호 pp. 30~37 (2022)
The Journal of the Acoustical Society of Korea Vol.41, No.1 (2022)
https://doi.org/10.7776/ASK.2022.41.1.030

pISSN : 1225-4428
eISSN : 2287-3775

주파수 영역 심층 신경망 기반 음성 향상을 위한 실수 네트워크와 복소 네트워크 성능 비교 평가

Performance comparison evaluation of real and complex networks for deep neural network-based speech enhancement in the frequency domain

황서림[1], 박성욱[2], 박영철[1†]

(Seo-Rim Hwang,[1] Sung Wook Park,[2] and Youngcheol Park[1†])

[1]연세대학교 지능형신호처리연구실, [2]강릉원주대학교 전자공학과

(Received December 3, 2021; revised December 27, 2021; accepted January 4, 2022)

초 록: 본 논문은 주파수 영역에서 심층 신경망 기반 음성 향상 모델 학습을 위하여 학습 대상과 네트워크 구조에 따라 두 가지 관점에서 성능을 비교 평가한다. 이때, 학습 대상으로는 스펙트럼 매핑과 Time-Frequency(T-F) 마스킹 기법을 사용하였고 네트워크 구조는 실수 네트워크와 복소 네트워크를 사용하였다. 음성 향상 모델의 성능은 데이터 셋 규모에 따라 Perceptual Evaluation of Speech Quality(PESQ)와 Short-Time Objective Intelligibility(STOI) 두 가지 객관적 평가지표를 통해 평가하였다. 실험 결과, 네트워크의 종류와 데이터 셋 종류에 따라 적정한 훈련 데이터의 크기가 다르다는 것을 확인하였다. 또한, 데이터의 크기와 학습 대상에 따라 복소 네트워크보다 실수 네트워크가 비교적 높은 성능을 보이기 때문에 총 파라미터의 수를 고려한다면 경우에 따라 실수 네트워크를 사용하는 것이 보다 현실적인 해결책일 수 있다는 것을 확인하였다.

핵심용어: 음성 향상, 스펙트럼 사상, T-F 마스킹, 복소 네트워크

ABSTRACT: This paper compares and evaluates model performance from two perspectives according to the learning target and network structure for training Deep Neural Network (DNN)-based speech enhancement models in the frequency domain. In this case, spectrum mapping and Time-Frequency (T-F) masking techniques were used as learning targets, and a real network and a complex network were used for the network structure. The performance of the speech enhancement model was evaluated through two objective evaluation metrics: Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI) depending on the scale of the dataset. Test results show the appropriate size of the training data differs depending on the type of networks and the type of dataset. In addition, they show that, in some cases, using a real network may be a more realistic solution if the number of total parameters is considered because the real network shows relatively higher performance than the complex network depending on the size of the data and the learning target.

Keywords: Speech enhancement, Spectral mapping, T-F masking, Complex network

PACS numbers: 43.60.Uv, 43.72.Ar

†**Corresponding author:** Youngcheol Park (young00@yon-sei.ac.kr)
Division of Computer and Telecommunication Engineering, Yonsei University, Chang jo room 269, 1 Yonseidae-gil, Wonju, Gangwon-do 26493, Republic of Korea
(Tel: 82-33-2744, Fax: 82-33-763-4323)

Copyright©2022 The Acoustical Society of Korea. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

I. 서 론

음성 향상은 잡음이 있는 음성 신호의 명료도와 품질을 향상시키는 기법으로 음성 통신, 보청기, 자동음성인식과 같이 소음 억제와 음성 특징 유지가 필수적인 분야에서 다양한 활용이 가능하다. 최근에는 심층 신경망(Deep Neural Network, DNN) 기반의 음성 향상 기법이 활발하게 연구되고 있으며, 기존의 확률 기반 기법과 비교하여 우수한 성능을 보이고 있다.[1-3]

심층 신경망 기반 음성 향상은 음성의 특징을 보다 효과적으로 학습하기 위하여 주로 시간 영역의 신호를 Short- Time Fourier Transform(STFT) 영역으로 변환하여 사용한다.[2,3] 이러한 주파수 영역 심층 신경망 기반 음성 향상을 설계하기 위한 학습 기법은 학습 대상과 네트워크 구조에 따라 크게 두 가지로 나눌 수 있다.

먼저, 심층 신경망 학습 대상에 따라 학습 기법은 스펙트럼 사상과 Time- Frequency(T-F) 마스킹 기법으로 나뉘는데, 스펙트럼 사상 기법은 잡음이 있는 음성으로부터 대상(깨끗한) 음성을 직접 추정하는 반면 T-F 마스킹 기법은 음성 고유의 특성을 유지하고 추정된 마스크를 잡음이 있는 음성에 곱하여 잡음을 억제한다. 이때, 스펙트럼 사상과 T-F 마스킹은 각각 음질과 명료도 개선에 좋은 성능을 보이는 것으로 알려져 있기 때문에 목적에 따라 적절한 학습 기법을 선정하는 것이 중요하다.[4]

두 번째로, 심층 신경망 네트워크 구조를 실수 네트워크로 할 것인지 복소 네트워크로 할 것인지에 따라 학습 기법을 나눌 수 있다. 초기의 주파수 영역 심층 신경망 기반 음성 향상 기법은 위상 정보 추정의 어려움으로 인해 크기 추정에 중점을 두었다.[1,2] 그리고 이렇게 추정한 크기 정보에 잡음이 포함된 입력 음성의 위상 정보를 결합하는 방법으로 최종 결과를 생성하였다. 그러나 이 경우, 특히 낮은 신호대잡음비(Signal-to-Noise Ratio, SNR) 상황에서 왜곡이 발생할 수 있다.[3,5] 이러한 왜곡의 발생을 막기 위해 위상 정보를 함께 추정할 수 있는 복소 스펙트럼 사상법[6]이나 다양한 복소 마스크를 추정하여 이용하는 방법이 사용되고 이를 위하여 기존의 실수 네트워크를 확장하여 복소수 스펙트럼이나 복소수 마스크의 실수부와 허수부를 모두 추정하는 복소 네트워크가[3,7] 제안되었다.

본 논문에서는 실수 네트워크와 복소 네트워크를 사용하여 스펙트럼 사상 기법과 T-F 마스킹 기법의 성능을 다양한 규모의 데이터 셋에서 비교 평가하고, 이를 통해 심층 신경망 기반 음성 향상에 효과적인 네트워크 구조와 데이터 셋 규모를 모색해 보았다.

II. 심층 신경망 기반 음성 향상 성능 평가를 위한 비교군

2.1 실수 네트워크와 복소 네트워크

시간 영역에서 잡음이 섞인 음성 신호를 y_t라고 할 때, y_t는 깨끗한 음성 신호 x_t와 잡음 신호 n_t의 합으로 표현 가능하다. 또한, y_t는 STFT을 통해 T-F 영역의 신호 $Y_{t,f}$로 변환할 수 있다. 이때, $Y_{t,f}$는 다음과 같이 표현할 수 있다.

$$|Y|e^{j\theta_Y} = |X|e^{j\theta_X} + |N|e^{j\theta_N}. \quad (1)$$

이때, $|\bullet|$와 θ, j는 각 성분의 크기와 위상, 허수 단위를 의미하며, 편의를 위하여 시간-주파수 첨자 t,f를 생략하였다.

초기 연구[1,2]에서는 신호 또는 T-F 마스크의 크기만을 추정하는 실수 네트워크를 사용하였다. 이때, 향상된 음성 $\hat{X}$을 합성할 때 잡음이 섞인 위상θ_Y을 그대로 사용하였는데 이러한 접근법은 성능 향상에 한계를 가지고 있다[3,5]. 최근 다양한 연구에서 이러한 문제를 해결하기 위해 복소 네트워크를 제안하였으며[3,7] 실수 네트워크와 비교하여 우수한 성능을 보였다. 복소 네트워크는 신호를 실수 부분 $\widehat{X_R}$과 허수 부분 $\hat{X}_I$를 모두 추정하므로, 이들을 이용하여 신호의 크기와 위상을 다음과 같이 얻을 수 있다.

$$|\hat{X}| = \sqrt{\widehat{X_R}^2 + \widehat{X_I}^2}. \quad (2)$$

$$\theta_{\hat{X}} = \tan^{-1}\frac{\hat{X}_I}{\hat{X}_R}. \quad (3)$$

본 논문은 실수 네트워크와 복소 네트워크를 다양한 환경에서 직접 비교하기 위하여, 최근 우수한 성능을 보인 복소 네트워크인 Deep Complex Convolutional Recurrent Network(DCCRN)와 DCCRN의 실수 네트워크(CRN) 간의 성능을 두 종류의 데이터 셋을 세 가지 규모로 나누어 비교 평가하였다.

2.2 스펙트럼 사상과 T-F 마스킹

T-F 영역에서 심층 신경망을 이용하여 음성 향상을 할 때는 주로 스펙트럼 사상과 T-F 마스킹 기법을 사용하여 최종 신호를 얻는다.[4] 이때, 각각 실수 네트워크를 사용하는지 복소 네트워크를 사용하는지에 따라 얻을 수 있는 최종 신호가 달라지는데, 실수 네트워크에서 스펙트럼 사상 기법을 사용하여 주파수 영역의 Y의 크기 $|Y|$를 X의 크기 $|X|$로 직접 사상하면 향상된 음성의 크기 $|\hat{X}|$를 구할 수 있다. 그러나 이 경우 2.1절에서 언급했듯이 θ_Y로 인한 왜곡이 발생한다.

Tan과 Wang[6] 등은 이러한 문제를 방지하기 위하여 Y_R과 Y_I로부터 $\hat{X}_R$과 $\hat{X}_I$을 직접 사상하는 복소 스펙트럼 사상을 제안하였다.

T-F 마스킹 기법은 음성을 직접 사상하는 스펙트럼 사상 기법과 달리 Y로부터 X를 합성하기 위한 마스크 M을 추정한다. 그리고 추정된 $\hat{M}$은 Y와 곱하여 $\hat{X}$을 만든다. 초기의 T-F 마스킹 기법은 스펙트럼 사상과 마찬가지로, $|Y|$로부터 $|M|$를 추정하여 $|\hat{X}|$를 구하고 θ_Y와 결합하여 $\hat{X}$를 합성하였다.

그러나, 이 경우에도 θ_Y로 인하여 원하지 않은 왜곡이 발생한다. 최근에는 이를 해결하기 위하여 복소 마스크나 복소 마스크를 효과적으로 계산할 수 있는 복소 네트워크가 제안되었는데, 대표적인 복소 마스크는 complex Ideal Ratio Mask(cIRM)[7]로 수식은 다음과 같다.

$$M_{cIRM} = \frac{Y_R X_R + Y_I X_I}{Y_R^2 + Y_I^2} + j\frac{Y_R X_I - Y_I X_R}{Y_R^2 + Y_I^2}. \quad (4)$$

이를 통해 $\hat{X}$는 다음과 같이 구할 수 있다.

$$\begin{aligned}\hat{X} &= \hat{M} * Y \quad (5)\\ &= (\hat{M}_R Y_R - \hat{M}_I Y_I) + j(\hat{M}_R Y_I + \hat{M}_I Y_R),\end{aligned}$$

이때, *는 복소 곱을 의미하며, $\hat{M}_R$과 $\hat{M}_I$는 각각 Eq. (4)을 통해 얻어진 마스크의 실수 부분과 허수 부분이다. 복소 네트워크는 $\hat{X}_R$과 $\hat{X}_I$ 또는 $\hat{M}_R$과 $\hat{M}_I$을 효과적으로 계산할 수 있게 설계된 네트워크이다.[3,7]

스펙트럼 사상과 T-F 마스킹 기법 모두 네트워크를 통해 구해진 $\hat{X}$는 최종적으로 Inverse STFT(ISTFT)을 통하여 시간 영역의 음성 신호 $\hat{x}$로 표현된다.

본 논문에서는 실수 네트워크를 위해 실수 스펙트럼 사상을 사용하였고 복소 네트워크를 위해 복소 스펙트럼 사상과 cIRM을 사용하였다.

III. 베이스라인 네트워크 및 실험 구성

3.1 베이스라인 네트워크

실험에 사용한 베이스라인 네트워크는 Fig. 1과 같다. 먼저, 복소 네트워크로는 DCCRN[Fig. 1(a)]을 사용하였으며, 실수 네트워크로는 CRN[Fig. 1(b)]을 사

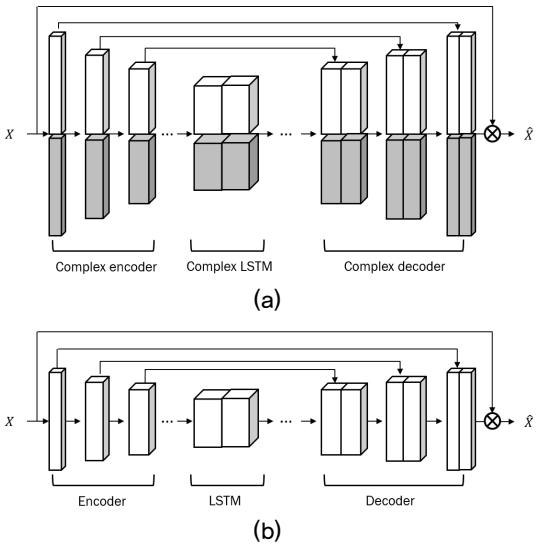

Fig. 1. The architecture of baseline network. (a) is complex network,[7] and (b) is real network of (a).

Table 1. CRN architecture. Here F denotes the number of frequency bins, and T denotes the number of time frames.

Layers	input size	parameters	output size
Conv2d	[1, F/2, T]	(5, 2), 16	[16, F/4, T]
Conv2d	[16, F/4, T]	(5, 2), 32	[32, F/8, T]
Conv2d	[32, F/8, T]	(5, 2), 64	[64, F/16, T]
Conv2d	[64, F/16, T]	(5, 2), 128	[128, F/32, T]
Conv2d	[128, F/32, T]	(5, 2), 128	[128, F/64, T]
Conv2d	[128, F/64, T]	(5, 2), 128	[128, F/128, T]
LSTM	[T, 128 * F/128]	256	[T, 128 * F/128]
ConvTrans2d	[256, F/128, T]	(5, 2), 128	[128, F/64, T]
ConvTrans2d	[256, F/64, T]	(5, 2), 128	[128, F/32, T]
ConvTrans2d	[128, F/32, T]	(5, 2), 128	[64, F/16, T]
ConvTrans2d	[64, F/16, T]	(5, 2), 64	[128, F/8, T]
ConvTrans2d	[32, F/8, T]	(5, 2), 32	[16, F/4, T]
ConvTrans2d	[16, F/4, T]	(5, 2), 16	[1, F/2, T]

용하였다.

DCCRN은 합성곱 계층으로 이루어진 인코더와 디코더 사이에 Long Short-Term Memory(LSTM) 계층을 갖는 네트워크이다. 이때, 각 계층은 입력의 실수 부분과 허수 부분을 위한 가중치를 별도로 갖으며 [Fig. 2(a)], 합성곱 계층에서는 복소 모듈[3]을 통하여 두 가중치를 복소 연산한다. CRN은 DCCRN에서 복소 연산을 구성하던 허수부 가중치와 복소 합성곱 모듈을 제거하여 사용하였다. DCCRN의 자세한 구조는 Reference [7]을 따랐으며 CRN 구조는 Table 1과 같다. Table 1에서 입출력 크기는 각각 [채널 수, 주파수 빈 수, 프레임 개수]를 나타낸 것이며, 파라미터는 Conv2d와 ConvTrans2d의 경우(커널 높이, 커널 폭)과 커널의 개수를 나타낸 것이고 LSTM의 경우 유닛의 개수를 나타낸 것이다.

윈도우 길이, 홉 길이, FFT는 각각 25 ms(T), 6.25 ms, 512 샘플(F)을 사용하였으며 DCCRN과 CRN의 총 파라미터 수는 각각 3.7 M,와 1.8 M이며 모델 최적화를 위해 Adam optimizer를 사용하였다. 손실함수로는 Mean Square Error(MSE)를 사용하였으며 학습률은 0.001로 정하였다.

3.2 실험 데이터 구성

실험에는 두 종류의 데이터 셋을 사용하였다. 먼저, 16 kHz로 샘플링된 TIMIT 음성 데이터 셋[10]을 사용하였으며, 잡음 데이터 셋은 NoiseX-92,[11] CHIME-2,[12] CHIME-3[13]를 혼합하여 사용하였다(dataset-1). 이때, 훈련 데이터 셋은 음성 데이터 셋 규모에 따라 세 가지로 나누었는데, TIMIT 음성 1,000개, 2,000개, 3,696개를 각각 랜덤으로 선택한 잡음 음성과 SNR 0 dB에서 20 dB까지 5 dB 간격으로 섞어 총 5,000개, 10,000개, 18,480개 세 가지 규모의 데이터 셋을 생성하였다. 테스트 데이터 셋은 훈련에 사용되지 않은 음성 462개를 훈련 데이터와 같은 방식으로 생성하였다.

또한, 데이터 셋에 따른 규칙성을 확인하기 위해 dataset-1보다 상대적으로 큰 규모인 DNS challenge 데이터 셋[9]의 일부를 사용하여 추가적인 실험을 진행하였다(dataset-2). dataset-2는 dataset-1과 동일한 방법으로 총 5,000개, 50,000개, 100,000개 세 가지 규모로 데이터 셋을 생성하였으며 462개의 테스트 데이터를 사용하였다. 이때, dataset-1과 dataset-2의 발화 길이는 동일하게 3초로 맞추어주었으며 dataset-1의 음성은 구어체와 문어체가 섞여있고 dataset-2는 책을 읽는 음성이다.

IV. 실험 결과 및 평가

모델 성능 평가는 각각 음질과 음성의 명료도 평가를 위해 가장 많이 사용되는 객관적 평가지표인 Perceptual Evaluation of Speech Quality(PESQ)[14]와 Short-Time Objective Intelligibility(STOI)[15]를 사용하였으며 실험 결과는 Tables 2~5와 같다. 각 테스트 별로 DCCRN과 CRN의 성능을 비교하여 더 높은 PESQ와 STOI 값을 굵은 글씨로 나타내었다.

먼저, Tables 2와 3은 데이터 셋 규모에 따른 스펙트럼 사상의 각 SNR 별 성능을 DCCRN과 CRN에서 평가한 결과이다. dataset-1의 경우(Table 2) 모든 SNR에서 DCCRN이 CRN보다 높은 STOI 값을 보였으며 10,000개 데이터를 사용했을 때를 제외하고 SNR 10 dB 이상일 때 DCCRN이 CRN보다 높은 PESQ 값을 보였다. 반면, dataset-2는(Table 3) 규모에 따라 성능 양상이 다르게 나타났는데, 5,000개 데이터를 사용했을 때 SNR 5 dB 이하인 경우 DCCRN의 PESQ와 STOI 값이 CRN보다 더 높았으며 10 dB 이상인 경우 CRN

Table 2. Performance evaluation of various network types using spectral mapping method on dataset-1.

SNR	Metric	Dataset scale	5,000		10,000		18,480	
		Noisy	CRN	DCCRN	CRN	DCCRN	CRN	DCCRN
0 dB	PESQ	1.15	**1.90**	**1.90**	**1.98**	**1.98**	**2.05**	2.04
	STOI	0.749	0.875	**0.887**	0.875	**0.900**	0.885	**0.907**
5 dB	PESQ	1.29	**2.33**	2.32	**2.46**	2.43	2.48	**2.50**
	STOI	0.833	0.925	**0.937**	0.927	**0.944**	0.930	**0.948**
10 dB	PESQ	1.58	2.84	**2.85**	**3.03**	2.96	2.96	**3.04**
	STOI	0.906	0.960	**0.965**	0.961	**0.969**	0.961	**0.972**
15 dB	PESQ	1.97	3.24	**3.29**	**3.44**	3.39	3.38	**3.46**
	STOI	0.950	0.977	**0.979**	0.979	**0.982**	0.981	**0.983**
20 dB	PESQ	2.51	3.56	**3.63**	**3.74**	3.70	3.66	**3.76**
	STOI	0.971	0.987	**0.988**	0.988	**0.989**	0.988	**0.990**

Table 3. Performance evaluation of various network types using spectral mapping method on dataset-2.

SNR	Metric	Dataset scale	5,000		50,000		100,000	
		Noisy	CRN	DCCRN	CRN	DCCRN	CRN	DCCRN
0 dB	PESQ	1.14	1.33	**1.36**	1.64	**1.65**	**1.66**	**1.66**
	STOI	0.755	0.789	**0.790**	**0.865**	0.860	0.867	**0.867**
5 dB	PESQ	1.21	1.51	**1.56**	1.97	**1.99**	1.97	**2.02**
	STOI	0.831	0.849	**0.852**	**0.913**	0.912	0.912	**0.913**
10 dB	PESQ	1.40	1.80	**1.81**	2.40	**2.43**	2.38	**2.47**
	STOI	0.893	**0.899**	0.892	**0.946**	0.944	**0.946**	0.945
15 dB	PESQ	1.69	**2.11**	2.05	2.72	**2.77**	2.70	**2.83**
	STOI	0.935	**0.923**	0.919	**0.963**	0.962	0.960	**0.961**
20 dB	PESQ	2.13	**2.33**	2.31	3.06	**3.14**	3.07	**3.18**
	STOI	0.965	**0.943**	0.940	**0.976**	0.975	0.976	**0.977**

Table 4. Performance evaluation of various network types using T-F masking method on dataset-1.

SNR	Metric	Dataset scale	5,000		10,000		18,480	
		Noisy	CRN	DCCRN	CRN	DCCRN	CRN	DCCRN
0 dB	PESQ	1.15	**1.92**	1.91	**1.97**	**1.97**	**2.06**	2.01
	STOI	0.749	0.868	**0.870**	0.882	**0.878**	0.880	**0.886**
5 dB	PESQ	1.29	**2.39**	2.37	2.41	**2.45**	**2.55**	2.51
	STOI	0.833	0.922	**0.926**	0.928	**0.930**	0.930	**0.936**
10 dB	PESQ	1.58	**2.95**	2.94	2.91	**2.99**	**3.07**	**3.07**
	STOI	0.906	0.958	**0.961**	0.960	**0.963**	0.963	**0.966**
15 dB	PESQ	1.97	3.37	**3.40**	3.30	**3.40**	**3.49**	**3.49**
	STOI	0.950	0.977	**0.979**	0.978	**0.980**	0.980	**0.982**
20 dB	PESQ	2.51	3.68	**3.72**	3.63	**3.70**	3.77	**3.78**
	STOI	0.971	0.987	**0.988**	0.988	0.989	0.989	**0.990**

이 더 높은 값을 보였다. 50,000개 데이터를 사용했을 때는 모든 SNR에서 PESQ 값은 DCCRN이 더 높았고 STOI 값은 CRN이 더 높았다. 그리고 100,000개 데이터를 사용했을 때는 SNR 10 dB를 제외하고 DCCRN의 PESQ와 STOI 값이 CRN보다 더 높았다.

Tables 4, 5는 데이터 셋 규모에 따른 T-F 마스킹의 각 SNR 별 성능을 DCCRN과 CRN에서 평가한 결과이다. dataset-1은(Table 4)의 경우 Table 2와 마찬가지로 모든 SNR에서 DCCRN이 CRN보다 높은 STOI 값을 가지며, 10,000개 데이터를 사용할 때 DCCRN이 모든 SNR에서 CRN보다 높은 PESQ 값을 갖는다. 반면, 5,000개 데이터를 사용할 때 SNR 10 dB 이하인 경

Table 5. Performance evaluation of various network types using T-F masking method on dataset-2.

SNR	Metric	Dataset scale	5,000		50,000		100,000	
		Noisy	CRN	DCCRN	CRN	DCCRN	CRN	DCCRN
0 dB	PESQ	1.14	1.35	**1.37**	**1.69**	1.67	**1.69**	1.68
	STOI	0.755	0.789	**0.794**	0.865	**0.865**	0.866	**0.868**
5 dB	PESQ	1.21	1.57	**1.57**	**2.04**	2.01	2.04	**2.07**
	STOI	0.831	0.852	**0.855**	0.914	**0.915**	0.914	**0.916**
10 dB	PESQ	1.40	**1.89**	1.86	**2.49**	2.46	**2.54**	2.54
	STOI	0.893	**0.901**	0.898	0.946	**0.948**	0.948	**0.950**
15 dB	PESQ	1.69	**2.20**	2.15	**2.87**	2.83	2.89	**2.92**
	STOI	0.935	**0.929**	0.924	**0.965**	0.965	0.964	**0.964**
20 dB	PESQ	2.13	**2.53**	2.43	**3.21**	3.19	3.27	**3.28**
	STOI	0.965	**0.948**	0.943	**0.977**	0.977	0.978	**0.978**

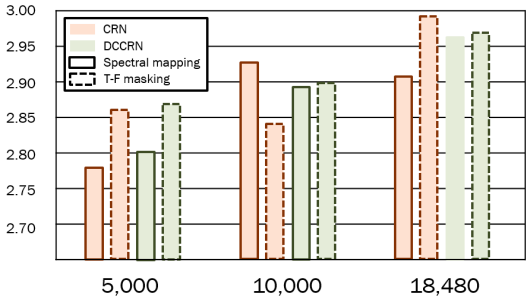

Fig. 2. (Color available online) Average PESQ results according to the scale of dataset-1.

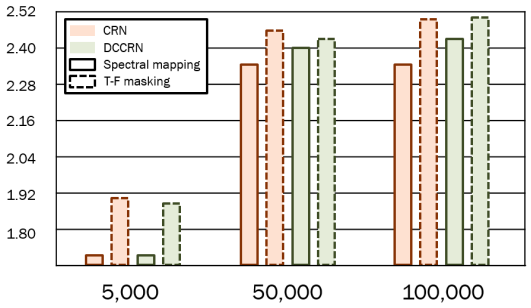

Fig. 3. (Color available online) Average PESQ results according to the scale of dataset-2.

우나 18,480개 데이터를 사용할 때 SNR 20 dB를 제외한 경우에는 CRN이 DCCRN보다 높은 PESQ 값을 갖는다. dataset-2의 경우(Table 5) Table 3과 유사한 경향성을 보이는데 10,000개 데이터를 사용했을 때 스펙트럼 사상의 경우 모든 SNR에서 DCCRN의 PESQ가 높았던 반면, T-F 마스킹의 경우 CRN이 높은 PESQ 값을 갖는다. 또한, dataset-2는 dataset-1과 비교하여 상대적으로 낮은 성능을 보였는데, 이는 dataset-2의 음성 대비 잡음 데이터비율이 상대적으로 더 높고 발화 길이를 맞춰주는 과정에서 묵음 구간이 많아져 모델 학습에 어려움을 겪었을 것으로 예상한다.

Tables 2 ~ 5를 볼 때, PESQ 값이 STOI 값에 비해 더 분명한 차이를 보였기 때문에 각 실험 결과에 따른 평균 PESQ를 데이터 셋 별로 Figs. 3과 4에 정리하였다. 이때, Figs. 2와 3은 각각 0.05와 0.12 길이로 스케일링하였다.

Figs. 2와3을 보면 DCCRN은 학습 기법에 따른 차이 없이 dataset-1과 dataset-2 모두 데이터 셋의 규모가 커질수록 PESQ 값이 향상되었다. 반면, CRN은 dataset-2에서는 DCCRN과 유사한 양상을 보였지만, dataset-1에서는 스펙트럼 사상을 사용한 경우 중간 규모의 데이터 셋이 가장 높은 PESQ 값을 보였고 T-F 마스킹을 사용한 경우 가장 낮은 PESQ 값을 보였다.

결과적으로 DCCRN과 CRN 사이에서 큰 성능 차이는 없었지만, dataset-1에서는 5,000개 데이터를 사용했을 때 T-F 마스킹을 사용하는 DCCRN의 성능이 가장 좋았으며, 10,000개와 18,480개 데이터를 사용했을 때 각각 스펙트럼 사상과 T-F 마스킹을 사용하는 CRN이 가장 좋은 성능을 보였다. dataset-2의 경우 5,000개와 50,000개 데이터를 사용했을 때 T-F 마스킹을 사용하는 CRN의 성능이 가장 좋았으며, 100,000개 데이터를 사용했을 때 T-F 마스킹을 사용하는 DCCRN의 가장 좋은 성능을 보였다.

한편, CRN은 DCCRN과 비교하여 사용하는 총 파라미터 수가 약 2배 가까이 적기 때문에 특정 상황에서는 CRN을 사용하는 것이 효율적일 수 있다.

V. 결 론

본 논문은 실수 네트워크와 복소 네트워크를 사용하여 스펙트럼 사상과 T-F 마스킹을 사용하였을 때 세 가지 규모의 데이터 셋 크기에 따른 성능 차이를 PESQ 와 STOI 관점에서 살펴보았다. 복소 네트워크의 경우 데이터 크기가 클수록 전체적인 성능이 개선되었지만, 실수 네트워크는 특히 높은 SNR 상황에서 중간 규모의 데이터 크기로 스펙트럼 사상을 사용하였을 때 더 좋은 성능을 보였다. 또한, 데이터 크기와 학습 기법에 따라 복소 네트워크 보다 실수 네트워크를 사용하는 것이 더 높은 성능을 보이기도 하였다. 따라서 총 파라미터의 수를 고려한다면 경우에 따라, 실수 네트워크를 사용하는 것이 보다 현실적인 해결책일 수 있다는 것을 확인하였다.

References

1. A. Narayanan and D. Wang, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” Proc. IEEE ICASSP. 7092-7096 (2013).
2. T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, “Phase processing for single-channel speech enhancement: History and recent advances,” IEEE Signal Process. Mag. **32**, 55-66 (2015).
3. H.-S. Choi, J-H Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee,“Phase-aware speech enhancement with deep complexu-net,” Proc. ICLR. 2019.
4. S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, “Mapping and masking targets comparison using different deep learning based speech enhancement architectures,” Proc. IJCNN. 1-8 (2020).
5. K. Paliwal, K. Wójcicki, and B. Shannon, “The importance of phase in speech enhancement,” Speech Commun. **53**, 465-494 (2011).
6. K. Tan and D. Wang, “Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement,” Proc. IEEE ICASSP. 6865-6869 (2019).
7. 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,” Proc. Interspeech, 2472-2476 (2020).
8. S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, “How does batch normalization help optimization?,” Proc. NeurIPS. 1-11 (2018).
9. C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dybey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, and J. Gehrke, “The interspeech 2020 deep noise suppression challenge: Dataset, subjective testing framework, and challenge results,” arXiv preprint arXiv:2005.13981 (2020).
10. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, “Timit acoustic phonetic continuous speech corpus,” Linguistic Data Consortium (1993).
11. A. Varga and H. J. M. Steeneken, “Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” Speech commun. **12**, 247-251 (1993).
12. E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, “The second ‘chime’speech separation and recog-nition challenge: Datasets, tasks and baselines,” Proc. IEEE ICASSP. 126-130 (2013).
13. J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘chime’ speech separation and recognition challenge: Dataset, taskand baselines,” Proc. ISRU. 504-511 (2015).
14. 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 coders,” Proc. IEEE ICASSP. 749-752 (2001).
15. C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” IEEE Trans. on Audio, Speech, and Lang. Process. **19**, 2125-2136 (2011).

저자 약력

▸ 황 서 림 (Seo-Rim Hwang)

2017년 3월 ~ 2021년 8월 : 연세대학교 컴퓨터정보통신공학부 학사 과정
2021년 9월 ~ 현재 : 연세대학교 일반대학원 전산학과 석사과정

▸ 박 성 욱 (Sung Wook Park)

1993년 2월 : 연세대학교 전자공학과 학사
1995년 2월 : 연세대학교 신호처리 석사
1998년 8월 : 연세대학교 신호처리 박사
2009년 3월 ~ 현재 : 국립강릉원주대학교 전자공학과 부교수
<관심분야> VLSI 신호처리, 멀티미디어 시스템

▸ **박 영 철 (Youngcheol Park)**

1986년 2월 : 연세대학교 전자공학과 학사
1988년 2월 : 연세대학교 전자공학과 석사
1993년 2월 : 연세대학교 전자공학과 박사
2002년 3월 ~ 현재 : 연세대학교 소프트웨어학부 교수