한국음향학회지 제43권 제2호 pp. 253~259 (2024)
The Journal of the Acoustical Society of Korea Vol.43, No.2 (2024)
https://doi.org/10.7776/ASK.2024.43.2.253

pISSN : 1225-4428
eISSN : 2287-3775

이중 분기 디코더를 사용하는 복소 중첩 U-Net 기반 음성 향상 모델

Complex nested U-Net-based speech enhancement model using a dual-branch decoder

황서림,[1] 박성욱,[2] 박영철[1†]

(Seorim Hwang,[1] Sung Wook Park,[2] and Youngcheol Park[1†])

[1]연세대학교 지능형신호처리연구실, [2]강릉원주대학교 전자공학과

(Received January 22, 2024; accepted February 7, 2024)

초 록: 본 논문에서는 이중 분기 디코더를 갖는 복소 중첩 U-Net 기반의 새로운 음성 향상 모델을 제안하였다. 제안된 모델은 음성 신호의 크기와 위상 성분을 동시에 추정할 수 있도록 복소 중첩 U-Net으로 구성되며, 디코더는 스펙트럼 사상과 시간 주파수 마스킹을 각각의 분기에서 수행하는 이중 분기 디코더 구조를 갖는다. 이때, 이중 분기 디코더 구조는 단일 디코더 구조에 비하여, 음성 정보의 손실을 최소화하면서 잡음을 효과적으로 제거할 수 있도록 한다. 실험은 음성 향상 모델 학습을 위해 보편적으로 사용되는 VoiceBank + DEMAND 데이터베이스 상에서 이루어졌으며, 다양한 객관적 평가 지표를 통해 평가되었다. 실험 결과, 이중 분기 디코더를 사용하는 복소 중첩 U-Net 기반 음성 향상 모델은 기존의 베이스라인과 비교하여 Perceptual Evaluation of Speech Quality(PESQ) 점수가 0.13가량 증가하였으며, 최근 제안된 음성 향상 모델들보다도 높은 객관적 평가 점수를 보였다.

핵심용어: 음성 향상, 복소 중첩 U-Net, 이중 분기 디코더, 스펙트럼 사상, 시간 주파수 마스킹

ABSTRACT: This paper proposes a new speech enhancement model based on a complex nested U-Net with a dual-branch decoder. The proposed model consists of a complex nested U-Net to simultaneously estimate the magnitude and phase components of the speech signal, and the decoder has a dual-branch decoder structure that performs spectral mapping and time-frequency masking in each branch. At this time, compared to the single-branch decoder structure, the dual-branch decoder structure allows noise to be effectively removed while minimizing the loss of speech information. The experiment was conducted on the VoiceBank + DEMAND database, commonly used for speech enhancement model training, and was evaluated through various objective evaluation metrics. As a result of the experiment, the complex nested U-Net-based speech enhancement model using a dual-branch decoder increased the Perceptual Evaluation of Speech Quality (PESQ) score by about 0.13 compared to the baseline, and showed a higher objective evaluation score than recently proposed speech enhancement models.

Keywords: Speech enhancement, Complex nested U-Net, Dual-branch decoder, Spectral mapping, Time-frequency masking

PACS numbers: 43.60.Uv, 43.72.Ar

I. 서 론

음성 향상은 다양한 배경 잡음으로부터 손상된 음성을 복원하는 기술로 음성 통신, 보청기, 자동음성 인식과 같이 의사 전달이 중요한 분야에서 필수적이다.[1] 기존의 음성 향상 기술은 Wiener filtering, spectral subtraction과 같은 기법을 사용하여 잡음을 제거하였는데, 이러한 확률통계 기반의 기법은 변화가 많은

†**Corresponding author:** Youngcheol Park (young00@yonsei.ac.kr)
Department of Software, Yonsei University, Chang jo room 265, 1 Yonseidae-gil, Wonju, Gangwon-do 26493, Republic of Korea
(Tel: 82-33-2744, Fax: 82-33-763-4323)

실제 환경 집음에 잘 대응하지 못한다는 한계를 지니고 있다. 최근 이러한 문제를 극복하기 위해 딥러닝 기반 음성 향상 기술이 주목받고 있다. 딥러닝을 이용한 접근법은 변화가 많은 잡음에도 잘 대응하며, 기존의 기법과 비교하여 우수한 성능을 보인다.[2,3]

딥러닝 기반 음성 향상 기술은 시간 영역과 시간-주파수 영역에서 적용될 수 있는데, 일반적으로 시간-주파수 영역에서 적용되었을 때 더 나은 성능을 제공하는 것으로 알려져 있다.[3] 시간-주파수 영역 음성 향상 기술은 모델 학습 대상에 따라 크게 스펙트럼 사상과 시간-주파수 마스킹 기법으로 나뉘며, 주로 스펙트럼 사상 기법은 음성 복원에, 시간-주파수 마스킹 기법은 잡음 제거에 탁월한 성능을 보인다.[4,5]

초기의 시간-주파수 영역 음성 향상 기술은 실수 영역에서 음성의 크기를 복원하고, 손상된 입력 음성의 위상을 재사용하는 방식으로 사용되었으나, 이 방법은 잡음이 섞인 위상 성분의 재사용으로 인한 음성 왜곡을 발생시킨다.[6] 이를 해결하기 위해 복소 영역에서 다양한 복소 마스크 추정 기법과 복소 네트워크가 제안되었고, 이는 일반적으로 기존의 실수 영역 기법보다 우수한 성능을 보인다.[5,7]

한편, 최근 제안된 중첩 U-Net(nested U-Net, same as U^2-Net) [7,8] 기반 음성 향상 모델은 인코더와 디코더의 각 계층을 U 모양의 블록으로 대체하여 우수한 성능을 보인 바 있다.

본 논문에서는 중첩 U-Net을 기본 구조로 하고, 위상 정보를 고려하여, 잡음 제거와 음성 복원을 동시에 성취할 수 있도록 개선된 음성 향상 모델을 제안하고자 한다. 이를 위하여 음성의 크기만 분석하고 복원하던 기본 구조를 복소수 연산이 가능하도록 개선하여 크기와 위상을 함께 분석할 수 있도록 만들고, 잡음 제거를 위주로 하는 디코더와 음성 복원을 위주로 하는 디코더를 결합한 이중 분기 디코더 구조[9,10]를 채택하였다. 그리고 다양한 평가 지표와 스팩트로그램을 사용하여 개선 정도를 확인하였다.

II. 기존 연구 내용

현재까지 많은 딥러닝 기반 음성 향상 모델들이 제안되었는데, 이러한 모델의 대부분이 인코더-디코더 구조를 사용하고 있다. 인코더-디코더 기반의 음성 향상 모델에서는 인코더에서 입력 음성을 압축하면서 잡음 제거와 동시에 음성에 대한 중요 특징을 추출하고, 추출된 특징을 디코더가 복원하여 최종 깨끗한 음성을 얻는다. 중첩 U-Net은 인코더-디코더 구조의 성능을 더 최적화하기 위하여 인코더와 디코더의 각 계층을 U 모양의 네트워크로 대체한 구조의 모델이다.

중첩 U-Net의 인코더와 디코더에서 사용되는 블록은 다음 식과 같이 표현 가능하다.

$$\chi_n = U(f(\chi_{n-1})) + f(\chi_{n-1}). \tag{1}$$

Eq. (1)에서 χ_n은 n 번째 블록의 출력을 의미하며, U는 인코더-디코더로 구성된 U 모양의 학습 가능한 복수 계층을, f는 학습 가능한 단층 계층을 나타낸다. 이때, 인코더에서는 다음 블록으로 넘어가기 전에 다운샘플링 과정을 거치며, 디코더에서는 U로 이전 블록의 출력이 전달되기 전에 업샘플링 과정을 거친다. 위 구조를 통해서 중첩 U-Net은 기존의 인코더-디코더 구조보다 더 다양한 스케일의 음성 특징을 추출할 수 있으며, 점진적인 다운샘플링과 업샘플링 과정을 통하여 인코딩 과정에서의 손실을 최소화할 수 있다.[7]

이중 분기 디코더는 기존의 대칭적으로 구성되었던 인코더-디코더 구조를 탈피하기 위한 시도 중의 하나이다. Reference [9]에서는 복소 스펙트럼을 추정하는 스펙트럼 디코더와 시간-주파수 마스크의 크기를 추정하는 시간-주파수 마스킹 디코더를 결합하였으며, 이후 Reference [10]에서는 복소 스펙트럼과 복소 시간-주파수 마스크를 추정하도록 하는 이중 분기 디코더를 제안하였다. 이를 수식으로 표현하면 다음과 같다.

$$\hat{X}_{t,f} = I(D_S(z), D_M(z)). \tag{2}$$

Eq. (2)에서 $\hat{X}_{t,f}$는 디코더를 통한 최종 출력을 나타내며, z는 인코더의 출력을 나타낸다. D_s와 D_M는 각각 스펙트럼 사상을 위한 디코더와 시간-주파수 마

스킹을 위한 디코더를 나타내며, $I(\cdot)$는 두 디코더를 결합하기 위해 사용되는 함수를 나타낸다.

Ⅲ. 제안하는 음성 향상 시스템

실수 영역 중첩 U-Net을 복소 영역으로 변환하기 위하여 학습 가능한 계층이 갖는 가중치(W)를 모두 실수 계수를 위한 가중치(W^R)와 허수 계수를 위한 가중치(W^I)로 분리하였다.[6] 그리고 각 계층에 대한 실수(O^R)와 허수 계수의 출력(O^I)을 Eqs. (3), (4)와 같이 얻어내었다. 이후 O^R와 O^I는 각각 정규화 함수와 활성화 함수를 통과한 뒤 다음 계층으로 이동한다.

$$O^R = W^R(u^I) - W^I(u^R). \tag{3}$$

$$O^I = W^R(u^R) + W^I(u^I). \tag{4}$$

3.1 이중 분기 복소 중첩 U-Net

잡음이 섞인 시간 영역의 음성 y_t가 입력되면 Short-Time Fourier Transform(STFT)을 통해 시간-주파수 영역에서의 $Y_{t,f} = Y_{t,f}^R + jY_{t,f}^I$를 얻는다. 이때, $Y_{t,f}^R$, $Y_{t,f}^I$는 각각 $Y_{t,f}$에 대한 크기와 실수, 허수 부분을 의미하며, t,f는 각각 시간과 주파수에 대한 첨자를 의미한다. $Y_{t,f}$는 모델의 입력으로 들어가 각각의 인코더-디코더를 통과하게 되는데, 이때 디코더는 Fig. 1과 같이 이중 분기 디코더를 사용하여 스펙트럼 추정과 시간-주파수 마스킹을 동시에 수행한다.

시간-주파수 마스킹을 위한 디코더의 출력을 $\hat{M}_{t,f}$, 스펙트럼 추정을 위한 디코더의 출력을 $\tilde{X}_{t,f}$라 할 때, 각 디코더 분기를 통해 나온 출력은 다음과 같이 주파수 영역에서 결합(Integration) 한다.

$$\overline{X}_{t,f} = \alpha |Y_{t,f}| |\hat{M}_{t,f}| \cdot e^{\theta_{Y_{t,f}} + \theta_{\hat{M}_{t,f}}} + \beta \tilde{X}_{t,f}. \tag{5}$$

Eq. (5)에서 α, β는 각 디코더 분기를 결합하기 위한 결합 계수이며, 실험을 통해 각각 1로 설정하였다. $|Y_{t,f}| = \sqrt{Y_{t,f}^{R2} + Y_{t,f}^{I2}}$와 $\theta_{Y_{t,f}} = \tan^{-1}(Y_{t,f}^I / Y_{t,f}^R)$는 각각 $Y_{t,f}$에 대한 크기와 위상을 나타낸다. 이때, 마스킹 디코더에서 향상된 음성을 구하는 방법($|Y_{t,f}| |\hat{M}_{t,f}| \cdot e^{\theta_{Y_{t,f}} + \theta_{\hat{M}_{t,f}}}$)은 Reference [11]에서 사용한 방법과 동일하다. 최종 구해진 $\overline{X}_{t,f}$는 inverse STFT(iSTFT)을 통해 시간 영역의 향상된 음성 $\overline{x}_{t,f}$로 변환한다.

이중 분기 중첩 U-Net은 기존의 단일 분기 디코더로 추정한 음성과 달리 서로 다르게 동작하는 두 개의 분기를 통해 추정한 음성을 결합한 것이기 때문

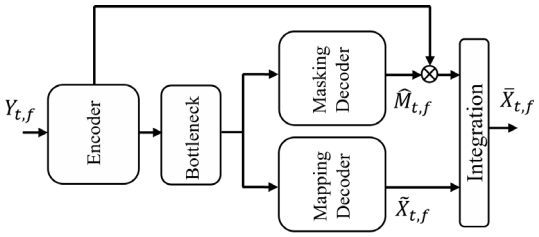

Fig. 1. Encoder-decoder architecture using a dual-branch decoder.

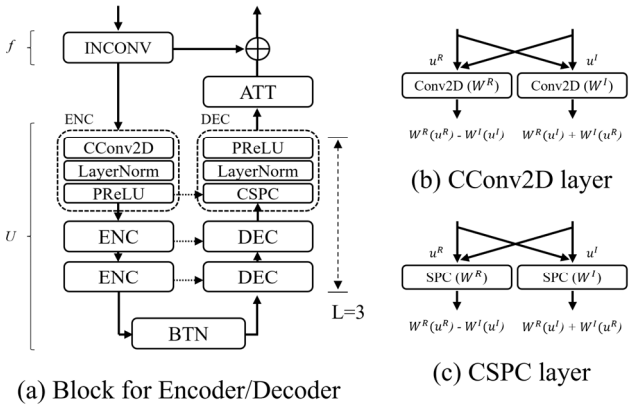

Fig. 2. Schemes of (a) Encoder/Decoder blocks, (b) CConv2D layer, and (c) CSPC layer.

에 각 분기 간의 장점을 효과적으로 반영할 수 있다.

제안된 이중 분기 디코더를 갖는 복소 중첩 U-Net의 인코더와 디코더는 각각 Fig. 2(a)에 나타난 블록을 기반하여 구성되어 있다. 그림에서는 편의를 위하여 다운샘플링 계층과 업샘플링 계층은 생략하였다. 이때, INCONV와 블록 내부 인코더(ENC), 블록 내부 병목 블럭(BTN), 블록 내부 디코더(DEC), 어텐션 모듈(ATT)은 각각 Eq. (1)의 f와 U에 대응된다. Fig. 2(b)와 (c)는 각각 Fig. 2(a)의 ENC와 DEC에 사용되는 복소 합성곱 계층을 나타내는데, 그림에서 나타나 있듯이, Eqs. (3)과 (4)를 통해 출력을 얻는다. INCONV, ATT, SPC는 각각 입력 채널을 증가시켜주는 합성곱 계층과 시간-주파수 어텐션 모듈,[8] 그리고 업샘플링 과정에 사용되는 sub-pixel 합성곱 계층[8]을 나타낸다.

3.2 손실함수

초기의 음성 향상 모델은 주로 평균 제곱 오차(Mean Squared Error, MSE) 함수를 사용하여 훈련되었다. 그러나 MSE는 모델의 성능을 최적화하는 데 있어 한계를 지니고 있으며,[12] 이를 보완하기 위해 다양한 손실함수가 제안되었다.[6,9] 압축된 주파수 결합 손실함수[9]는 최근 제안된 손실함수 중 가장 우수한 성능을 보이는 손실함수 중 하나이다.

본 논문에서도 음성 향상 모델 최적화를 위해 압축된 주파수 결합 손실 함수(L)를 사용하였으며 수식은 다음과 같다.

$$L = \gamma_1 L_m + \gamma_2 L_c. \tag{6}$$

$$L_m = \sum_{t,f} \| \, |\overline{X}_{t,f}| - |X_{t,f}|^c \, \|_2. \tag{7}$$

$$L_c = \sum_{t,f} \| \overline{X}^R_{t,f} - X^{R^c}_{t,f} \|_2 + \sum_{t,f} \| \overline{X}^I_{t,f} - X^{I^c}_{t,f} \|_2. \tag{8}$$

Eq. (6)에서 γ_1, γ_2는 각각 크기 손실함수(L_m)와 복소 손실 함수(L_c)를 위한 결합 계수를 의미하며 본 논문에서는 0.9와 0.1을 사용하였다. $\| \cdot \|_2$은 L2 norm을 의미하며, $X_{t,f}$는 주파수 영역에서의 깨끗한 타겟 음성을 의미한다. c는 압축을 위한 계수이며 Reference [9]에서는 0.5를 사용하였다. 본 논문에서는 0.1 ~ 0.5까지 0.1 단위로 값을 높여가며 실험하였으며, 실험 결과를 기반으로 0.2를 사용하였다.

위 손실함수를 통해 음성 향상 모델은 주파수 영역에서 추정 음성과 타겟 음성 간의 크기 차이와 실수, 허수 부분의 차이를 최소화하는 방향으로 학습하는 것이 가능하다.

IV. 실험 환경

실험을 위한 데이터베이스로는 음성 향상에서 주로 사용되는 VoiceBank + DEMAND(VBD) 데이터베이스[13]를 사용하였다. VBD는 11,572 개의 훈련 데이터와 824개의 테스트 데이터로 구성되어 있다. 이때, 훈련 데이터는 28명의 영어권 화자를 통해 녹음된 영어 발화를 각각 0 dB, 5 dB, 10 dB, 15 dB 신호 대 잡음비(Signal-to- Noise Ratio, SNR)로 음성과 잡음 신호를 섞어서 생성한 데이터이며, 테스트 데이터는 훈련에 사용되지 않은 2명의 영어권 화자를 통해 녹음된 영어 발화를 각각 2.5 dB, 7.5 dB, 12.5 dB, 17.5 dB SNR로 음성과 잡음 신호를 섞어서 생성한 데이터이다.

베이스라인 모델로는 Reference [8]에서 제안된 'Baseline + CTFA' 모델을 약간 수정하여 사용하였다. 이때, 수정한 부분은 다음과 같다. 1) 어텐션 모듈의 커널 크기를 1에서 17로 조정. 2) 블록 외부 병목 블록의 마지막 PReLU 활성화 함수 채널 수를 1에서 64로 조정. 3) ENC의 잔차 경로 제거.

실험에 사용한 모든 음성 및 잡음 데이터는 16 kHz로 샘플링 하였다. 윈도우 길이, 홉 길이, FFT 빈 개수, 청크 길이는 각각 25 ms, 6.25 ms, 512 샘플, 2 s를 사용하였다. 모델 훈련을 위해서는 Adam optimizer를 사용하였으며 배치 크기는 2를 사용하였다.

성능 평가는 Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility(STOI) 점수와 composite signal intelligibility(CSIG), composite background noise(CBAK), composite overall quality(COVL)를 포함한 3가지 OVL 점수[14]를 사용하였다. PESQ는 4.5, STOI는 %로 표시하여 100점 만점이며, OVL 점수는 5점 만점이다. 4가지 평가 지표 모두 높을수록 좋은 성능을 의미한다.

V. 실험 결과 및 분석

먼저, 제안된 복소 중첩 U-Net과 이중 분기 디코더의 성능을 확인하기 위해 실험(Ablation test)을 진행하였다. 결과는 Table 1에 나타나 있다. 이때, 베이스라인 모델(C: ✗, DB: ✗)은 시간-주파수 마스크를 타겟으로 한다. 실험 결과, 실수 영역 중첩 U-Net을 복소 영역으로 변환하였을 때(C: ✓, DB: ✗), CSIG 점수와 COVL 점수는 다소 하락했지만, CBAK 점수와 STOI 점수가 향상되었으며, 결과적으로 PESQ 점수가 0.09 가량 크게 증가한 것을 확인할 수 있다. 이 결과는 잡음 제거 성능이 높아지면서 전반적인 음성의 퀄리티가 높아졌음을 보여준다.

중첩 U-Net을 복소 영역으로 변환하고 이중 분기 디코더를 사용하였을 경우(C: ✓, DB: ✓), 이중 분기 디코더를 사용하지 않았을 때와 비교할 때, PESQ 점수가 0.04가량 증가하였으며, 배경 잡음에 대한 CBAK 점수는 3.66으로 같지만, 음성의 명료도를 나타내는 CSIG 점수가 4.25에서 4.39로 크게 개선된 것을 확인할 수 있다. 이는 스펙트럼 사상 디코더가 마스킹 과정에서 손실된 음성을 잘 복원시켜줌을 의미하며 마스킹 디코더와 스펙트럼 사상 디코더를 결합한 이중 분기 디코더가 서로 상호보완적으로 잘 동작함을 나타낸다. 이를 통해, 결과적으로 이중 분기 디코더를 사용하였을 때, 음성의 전체적인 점수를 나타내는 COVL 점수 또한 3.73에서 3.83으로 크게 증가하였다.

한편, 제안된 모델은 기존의 베이스라인과 비교하여 약 84 %의 학습 파라미터만으로 훨씬 우수한 점수를 보였다. 이는 제안 모델의 성능 향상이 단순히 파라미터의 증가로 인한 것이 아님을 보여준다. 이때, 제안된 모델이 베이스라인 모델(3.51 M)과 비교하여 상대적으로 낮은 파라미터(2.09 M)를 유지하는 이유는 베이스라인 모델을 복소 영역으로 변환하면서 합성곱 계층의 입출력 채널 개수를 고정한 채로 실수 부분과 허수 부분으로 나누었기 때문이다.

다음으로 최근 제안된 음성 향상 모델과의 비교를 진행하였다. 비교 평가에는 중첩 U-Net 기반의 음성 향상 모델인 GaGNet,[15] NUNet-TLS[8]를 포함하여 총 5개의 모델을 사용하였다. 이때, 모든 모델은 causality를 만족하며 결과는 Table 2에 나타나 있다. 해당 실험 결과는 각 논문에 나타난 수치를 옮겨 적은 값이며, 논문에서 제공되지 않은 값은 '-'으로 표시하였다.

Table 1. Ablation test for complex nested U-Net using a dual-branch decoder.

C	DB	Param.	Metric				
			PESQ	CSIG	CBAK	COVL	STOI
✗	✗	3.51 M	3.07	4.40	3.60	3.76	94.76
✓	✗	2.09 M	3.16	4.25	3.66	3.73	94.92
✓	✓	2.98 M	**3.20**	**4.39**	**3.66**	**3.83**	**95.00**

Table 2. Performance comparison with recent proposed speech enhancement models. All systems in this table satisfy causality.

Model	Param.	Metric				
		PESQ	CSIG	CBAK	COVL	STOI
Noisy	-	1.97	3.35	2.44	2.63	92.10
GaGNet[15]	5.94 M	2.94	4.26	3.45	3.59	94.70
DEMUCS[16]	128 M	3.07	4.31	3.40	3.63	95.00
NUNet-TLS[8]	2.83 M	3.04	4.38	3.47	3.74	94.76
CTS-Net[17]	4.35 M	2.92	4.25	3.46	3.59	-
FRCRN[18]	6.9 M	3.21	4.23	3.64	3.73	-
Proposed	**2.98 M**	**3.20**	**4.39**	**3.66**	**3.83**	**95.00**

실험 결과, 제안된 모델은 대부분의 평가 지표에서 가장 높은 점수를 보였다. 이를 통해 우리의 제안 모델이 기존의 중첩 U-Net의 성능을 높인 것뿐만 아니라, 다른 최신의 모델과 비교하여서도 우세한 성능을 보임을 알 수 있다. 한편, 제안 모델은 NUNet-TLS (스펙트럼 사상을 사용하는 중첩 U-Net 기반 모델)와 비슷한 CSIG 점수를 보이지만, 상대적으로 높은 CBAK 점수(NUNet-TLS: 3.47, Proposed: 3.66)와 COVL 점수(NUNet-TLS: 3.74, Proposed: 3.83)를 보인다. 이는 이중 분기 디코더의 사용으로 인한 것으로 추정한다. 한편, 제안된 모델은 FRCRN[18]보다 PESQ 측면에서 0.01 정도 낮은 수치를 보이지만, FRCRN과 비교하여 약 43 %의 학습 파라미터만으로 상대적으로 높은 CSIG, CBAK, COVL 점수를 보인다.

마지막으로 이중 분기 디코더의 효과를 확인하기 위해, 깨끗한 음성[Fig. 3(a)]과 이중 분기 디코더에서 스펙트럼 추정 분기의 출력을 통한 음성[Fig. 3(b)], 시간-주파수 마스킹 분기의 출력을 통한 음성[Fig. 3(c)], 두 분기의 출력을 결합하여 만들어진 최종 향

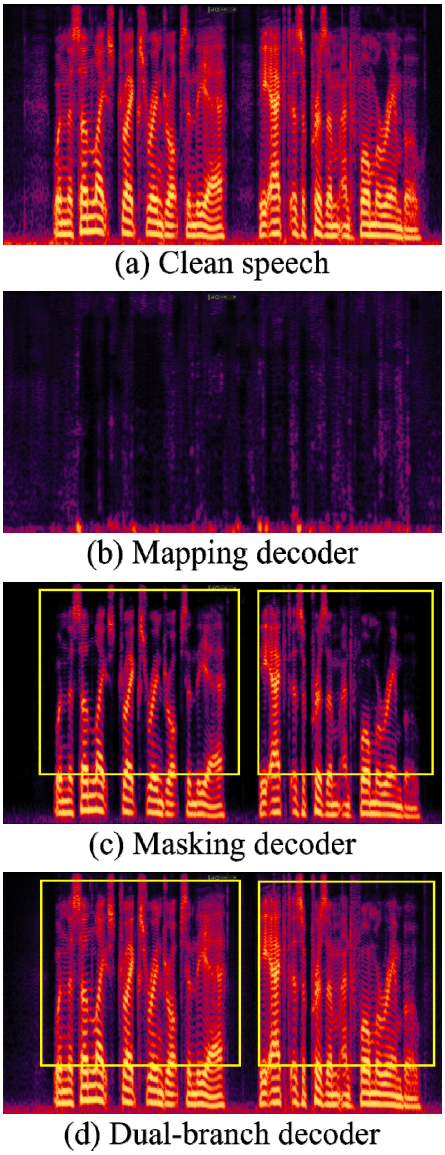

Fig. 3. (Color available online) The spectrogram of (a) clean speech, output speech of (b) mapping decoder, (c) masking decoder, and (d) dual-branch decoder.

상된 음성[Fig. 3(d)]을 각각 스펙트럼상에서 나타내었다. 이때, 관찰을 용이하게 하기 위해 스펙트럼 추정 분기의 출력은 20 dB 증폭하여 나타내었으며, 해당 음성은 824개의 테스트 데이터 음성중 하나를 랜덤하게 선별하였다.

Fig. 3(c)의 노란색 박스를 보면, 시간-주파수 마스킹 디코더의 출력에서는 잡음은 많이 제거되는 만큼 음성 정보 또한 많이 손실된 것을 확인할 수 있다. 반면, 이중 분기 디코더의 출력에서는 잡음이 많이 제거된 상태에서 음성 요소 또한 제대로 복원된 것을 확인할 수 있다. 이는 이중 분기 디코더에서 스펙트럼 사상과 시간-주파수 마스킹 분기가 상호 작용하여 동작하고 있음을 나타내며, 높은 CBAK 점수를 유지하면서 CSIG 점수가 크게 향상되었던 결과 분석과도 일치한다.

VI. 결 론

본 논문에서는 복소 영역에서 동작하는 중첩 U-Net을 기반으로 스펙트럼 사상과 시간-주파수 마스킹을 독립적으로 수행하는 이중 분기 디코더 구조를 갖는 음성 향상 모델을 제안하였다. 실험 결과, 입력을 복소 영역에서 처리함으로써 기존 중첩 U-Net 모델의 성능이 크게 향상되었다. 또한, 디코더를 이중으로 분기함으로써 추가적으로 성능을 개선할 수 있었다. 각 분기 출력 간의 스펙트럼 비교 분석을 통해, 시간-주파수 마스킹 디코더는 음성을 손실하더라도 최대한 많은 잡음을 제거하기 위해 노력하며, 스펙트럼 사상 디코더는 시간-주파수 마스킹 디코더가 손실한 음성을 복원하고자 노력하는 것을 확인하였다. 제안된 모델은 최신 음성 향상 모델과 비교하여서도 매우 우수한 성능을 보였다.

References

1. P. C. Loizou, *Speech Enhancement: Theory and Practice, 2nd ed.* (CRC Press, Inc., Boca Raton, 2013), pp. 1-768.
2. S. Pascual, A. Bonafonte, and J. Serra, “SEGAN: Speech enhancement generative adversarial network,” arXiv preprint arXiv:1703.09452 (2017).
3. H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, A. Ju, M. Zohourian, M. Tang, H. Gamper, M. Golestaneh, and R. Aichner, “Icassp 2023 deep speech enhancement challenge,” arXiv preprint arXiv: 2303.11510 (2023).
4. S. Hwang, S. W. Park, and Y. Park. “Performance comparison evaluation of real and complex networks for deep neural network-based speech enhancement in the frequency domain” (in Korean), J. Acoust. Soc. Kr. **41**, 30-37 (2022).
5. 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, IEEE IJCNN, 1-8 (2020).
6. 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," arXiv preprint arXiv:1903.03107 (2019).
7. X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested U-structure for salient object detection," Pattern Recognition, **106**, 107404 (2020).
8. S. Hwang, S. W. Park, and Y. Park, "Monoaural speech enhancement using a nested U-net with two-level skip connections," Proc. Interspeech, 191-195. (2022).
9. R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based metric GAN for speech enhancement," arXiv preprint arXiv:2203.15149 (2022).
10. Z. Zhang, S. Xu, X. Zhuang, L. Zhou, H. Li, and M. Wang, "Two-stage UNet with multi-axis gated multilayer perceptron for monaural noisy-reverberant speech enhancement," Proc. IEEE ICASSP, 1-5 (2023).
11. 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).
12. S. Hwang, J. Byun, and Y.-C. Park. "Performance comparison evaluation of speech enhancement using various loss functions" (in Korean), J. Acoust. Soc. Kr. **40**, 176-182 (2021).
13. C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," Proc. SSW, 146-152 (2016).
14. Y. Hu and P. C. Loizou. "Evaluation of objective measures for speech enhancement," Proc. Interspeech, 1447-1450 (2006).
15. A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," Appl. Acoust. **187**, 108499 (2022).
16. A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," arXiv preprint arXiv:2006.12847 (2020).
17. A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," IEEE/ACM Transa. on Audio, Speech, and Lang. Process. **29**, 1829-1843 (2021).
18. S. Zhao, B. Ma, K. N. Watcharasupat, and W. S. Gan, "FRCRN: Boosting feature representation using frequency recurrence for monaural speech enhancement," Proc. IEEE ICASSP, 9281-9285 (2022).

저자 약력

▸ 황 서 림 (Seorim Hwang)

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

▸ 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

▸ 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