특집논문 (Special Paper)
방송공학회논문지 제27권 제6호, 2022년 11월 (JBE Vol.27, No.6, November 2022)
https://doi.org/10.5909/JBE.2022.27.6.840
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

중첩 U-Net 기반 음성 향상을 위한 다중 레벨 Skip Connection

황 서 림[a)], 변 준[a)], 허 준 영[a)], 차 재 빈[b)], 박 영 철[c)‡]

Multi-level Skip Connection for Nested U-Net-based Speech Enhancement

Seorim Hwang[a)], Joon Byun[a)], Junyeong Heo[a)], Jaebin Cha[b)], and Youngcheol Park[c)‡]

요 약

심층 신경망(Deep Neural Network) 기반 음성 향상에서 입력 음성의 글로벌 정보와 로컬 정보를 활용하는 것은 모델의 성능과 밀접한 연관성을 갖는다. 최근에는 다중 스케일을 사용하여 입력 데이터의 글로벌 정보와 로컬 정보를 활용하는 중첩 U-Net 구조가 제안되었으며, 이러한 중첩 U-Net은 음성 향상 분야에도 적용되어 매우 우수한 성능을 보였다. 그러나 중첩 U-Net에서 사용되는 단일 skip connection은 중첩된 구조에 알맞게 변형되어야 할 필요성이 있다. 본 논문은 중첩 U-Net 기반 음성 향상 알고리즘의 성능을 최적화하기 위하여 다중 레벨 skip connection(multi-level skip connection, MLS)을 제안하였다. 실험 결과, 제안된 MLS는 기존의 skip connection과 비교하여 다양한 객관적 평가 지표에서 큰 성능 향상을 보이며 이를 통해 MLS가 중첩 U-Net 기반 음성 향상 알고리즘의 성능을 최적화시킬 수 있음을 확인하였다. 또한, 최종 제안 모델은 다른 심층 신경망 기반 음성 향상 모델과 비교하여서도 매우 우수한 성능을 보인다.

Abstract

In a deep neural network (DNN)-based speech enhancement, using global and local input speech information is closely related to model performance. Recently, a nested U-Net structure that utilizes global and local input data information using multi-scale has bee n proposed. This nested U-Net was also applied to speech enhancement and showed outstanding performance. However, a single skip connection used in nested U-Nets must be modified for the nested structure. In this paper, we propose a multi-level skip connection (MLS) to optimize the performance of the nested U-Net-based speech enhancement algorithm. As a result, the proposed MLS showed excellent performance improvement in various objective evaluation metrics compared to the standard skip connection, which means th at the MLS can optimize the performance of the nested U-Net-based speech enhancement algorithm. In addition, the final proposed m odel showed superior performance compared to other DNN-based speech enhancement models.

Keyword : Speech enhancement, multi-scale, nested U-Net, skip connection

a) 연세대학교 일반대학원 전산학과(The department of computer science, Yonsei University)
b) 연세대학교 컴퓨터정보통신공학부(The division of computer and telecommunications engineering, Yonsei University)
c) 연세대학교 소프트웨어학부(The division of software, Yonsei University)
‡ Corresponding Author : 박영철(Youngcheol Park)
E-mail: young00@yonsei.ac.kr
Tel: +82-33-760-2756
ORCID: https://orcid.org/0000-0003-3274-076X
· Manuscript August 1, 2022; Revised September 26, 2022; Accepted September 26, 2022.

Ⅰ. 서 론

음성 향상은 불필요한 배경 잡음을 제거하여 깨끗한 음성을 복원해내는 기술로 음성 인식 인공지능과 보청기, 야외 영상 촬영 등 음질과 음성의 명료도가 중요한 응용 분야에 필수적으로 사용되고 있다. 기존에는 확률 통계 기반의 기법을 사용하여 잡음을 제거하였는데, 딥러닝이 발전하면서 심층 신경망(Deep Neural Network) 기반 음성 향상 기술이 다양하게 연구되고 있으며 기존의 확률 통계 기반 기법과 비교하여 매우 우수한 성능을 보인다[1-4]. 이때, 심층 신경망 기반 음성 향상 기술은 주로 입력 음성 신호를 시간-주파수 영역으로 변환하여 사용한다[5].

최근에는 심층 신경망 모델이 음성을 더 잘 분석할 수 있도록 음성의 글로벌 정보와 로컬 정보를 활용하는 시도가 늘고 있으며[3,4,6], 이러한 노력의 일환으로 다중 스케일 특징 맵이 사용되고 있다. 참고문헌[6]은 합성곱 계층의 커널 크기를 다양하게 사용하여 다중 스케일 특징 맵을 추출하였고, 참고문헌[3]은 보조 네트워크를 사용하여 다중 스케일 특징 맵을 추출하였다. 그러나 두 경우 모두 모델의 파라미터 수가 많이 증가하고 계신 복잡도가 너무 커질 수 있다는 단점이 있다.

이러한 단점을 해결하기 위해 계산 복잡도 증가를 최소화하면서 다중 스케일 특징 맵을 사용할 수 있는 중첩 U-Net(Nested U-Net) 구조가 이미지 처리 분야에서 처음 제안되었다[1]. 이때, 중첩 U-Net은 U-Net의 각 계층을 U 모양의 작은 잔차 블록으로 대체하여 각 계층이 각기 다른 크기의 수용 영역(receptive field)을 사용할 수 있게 하였으며, U 모양의 잔차 블록의 파라미터 수를 기존 계층의 파라미터 수보다 적게 유지하여 큰 계산의 증가 없이 모델 깊이를 증가시키고 효과적으로 다중 스케일 특징 맵을 추출한다. 중첩 U-Net의 이러한 장점은 음성 향상 분야에서도 동일하게 적용됨이 증명되었다[4]. 그러나 중첩 U-Net[1,4]은 일반적인 U-Net과 같이 단일 skip connection을 사용한다.

Skip connection은 U-Net 기반 모델의 성능과 밀접한 연관성을 가진다[7]. 본 논문은 중첩 U-Net 기반 음성 향상 알고리즘의 성능을 최적화하기 위하여 중첩 구조에 적합한 다중 레벨 skip connection(multi-level skip connection, MLS)을 제안하였다. 이를 위해 일반적인 단일 skip connection에 추가적인 연결 경로를 다양하게 조합하고 결합하여 성능을 비교 평가하였다. 최종 제안된 MLS는 기존의 단일 skip connection보다 중첩 U-Net 기반 음성 향상 알고리즘의 성능을 최적화하는데 탁월한 능력을 보였으며 다른 심층 신경망 기반 음성 향상 모델과 비교하여서도 매우 우수한 성능을 보였다.

Ⅱ. 심층 신경망 기반 음성 향상 알고리즘과 중첩 U-Net

1. 시간-주파수 영역에서의 심층 신경망 기반 음성 향상 기법

심층 신경망 기반 음성 향상은 음성을 어떤 영역에서 처리하는지에 따라 크게 시간 영역 기법과 시간-주파수 영역 기법 두 가지로 나눌 수 있다. 이때, 시간-주파수 영역 기법은 시간 영역의 음성을 시간-주파수 영역으로 변환하여 잡음 제거를 위한 시간-주파수 마스크를 추정하거나 깨끗한 음성을 직접 매핑하며[8], 시간 영역 기법과 비교하여 음성의 특징을 다루는데 더 용이하다[5].

잡음이 섞인 시간 영역의 음성 신호 y_t는 깨끗한 음성 신호 x_t와 잡음 신호 n_t를 가산하여 만든다. 이때, 심층 신경망의 입력으로 y_t가 들어오면 y_t는 Short-Time Fourier Transform(STFT)을 통해 다음과 같이 시간-주파수 영역으로 변환된다.

$$|Y_{t,f}|e^{j\theta_{Y_{t,f}}} = |X_{t,f}|e^{j\theta_{X_{t,f}}} + |N_{t,f}|e^{j\theta_{N_{t,f}}} \quad (1)$$

위 식에서 $|\bullet|$와 θ, j는 각 성분의 크기와 위상, 허수 단위를 나타낸다.

본 논문에서는 $|Y_{t,f}|$로부터 깨끗한 음성의 크기를 직접 매핑하며 다음과 같이 향상된 음성 신호 $\hat{x}_t$을 구한다.

$$|\hat{X}_{t,f}| = DNN(|Y_{t,f}|) \quad (2)$$

$$\hat{X}_{t,f} = |\hat{X}_{t,f}|e^{j\theta_{Y_{t,f}}} \quad (3)$$

$$\hat{x}_t = ISTFT(\hat{X}_{t,f}) \quad (4)$$

위 식에서 DNN은 실험에 사용된 심층 신경망(중첩 U-Net)을 의미하며 $ISTFT$은 Inverse STFT(ISTFT)를 의미한다.

2. 중첩 U-Net 구조 기반의 음성 향상

중첩 U-Net은 중요 객체 검출 분야에서 처음 제안된 구조로[1] 이후 음성 향상 분야에도 적용되어 매우 우수한 성적을 보였다[4]. 이때, 중첩 U-Net은 차원 축소를 통해 작은 크기의 커널을 사용하여도 넓은 범위의 수용 영역을 가질 수 있으며 효과적으로 다중 스케일 특징 맵을 추출할 수 있다. 또한, 중첩 U-Net은 점진적으로 다운/업 샘플링을 진행하기 때문에 일반적인 U-Net과 비교하여 데이터의 의미적 차이(semantic gap)를 줄일 수 있다는 장점이 있다.

중첩 U-Net은 크게 합성곱으로 구성된 입출력 계층과 인코더-디코더 단계, 병목 블록으로 구성되어 있으며 입력이 들어오면 입력 계층을 통해 데이터의 특징 맵을 추출하고 차례대로 인코더 단계, 병목 블록, 디코더 단계를 거쳐 원하는 특징 맵을 구한다. 그리고 마지막 출력 계층을 통해 특징 맵의 차원 수를 입력과 동일하게 맞춰준다. 각각의 인코더-디코더 단계는 U 모양의 잔차 블록으로 이루어져 있으며, 인코더 단계의 출력단과 디코더 단계의 입력단에는 각각 다운 샘플링 계층과 업 샘플링 계층이 있다. 자세한 구조는 그림 1에서 확인할 수 있다.

Ⅲ. 다중 레벨 skip connection을 사용하는 중첩 U-Net

일반적인 U-Net과 같이 중첩 U-Net은 디코더 단계의 업

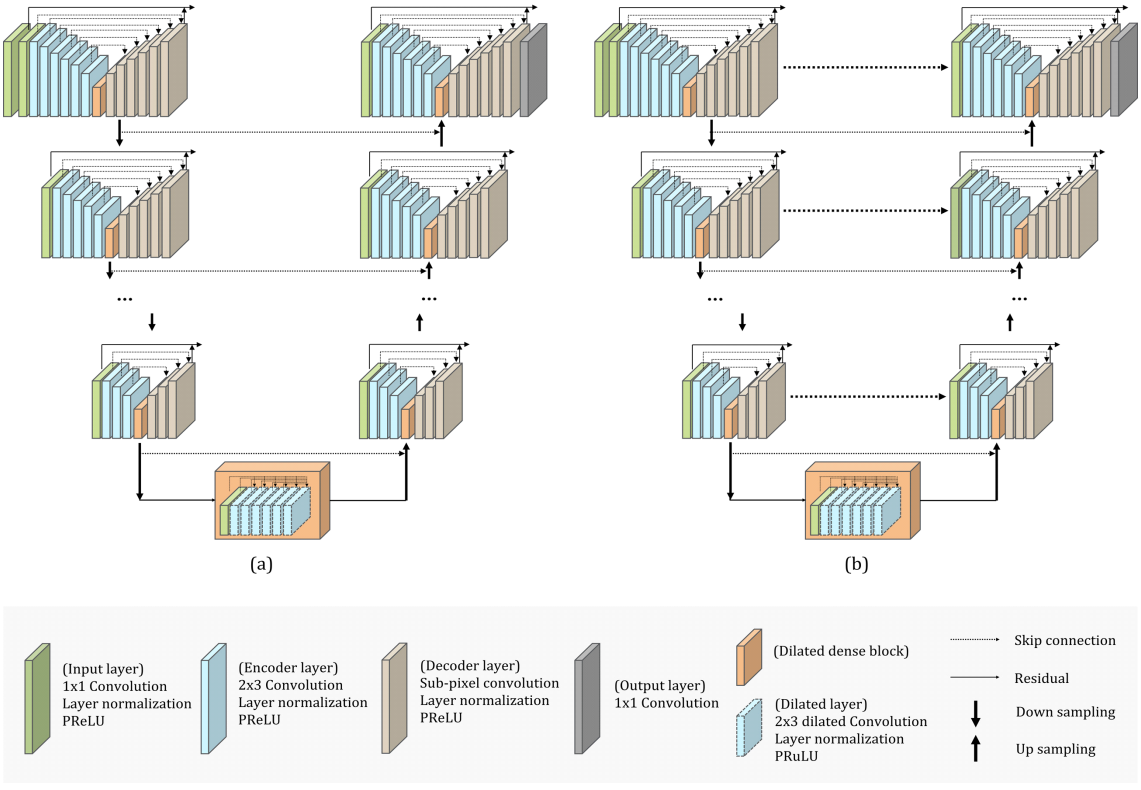


그림 1. 중첩 U-Net 기반 모델 구조. (a)는 추가적인 skip connection이 없는 경우의 구조를 나타내며, (b)는 논문에서 제안된 skip connection을 포함한 경우의 구조를 나타낸다

Fig. 1. The architecture of nested U-Net. (a) is without additional skip connection, and (b) is with the additional skip connection proposed in this paper

샘플링 과정에서 자세한 정보를 보완해주기 위하여 인코더 단계의 출력값을 디코더 단계의 입력으로 전달한다. 본 논문은 이러한 연결을 단일 skip connection이라 하였다. n번째 디코더 단계의 출력을 $\hat{X}_n^d$, 이에 대칭되는 $\bar{n}$번째 인코더 단계의 출력을 $\hat{X}_{\bar{n}}^e$이라고 할 때, 단일 skip connection은 다음과 같이 표현할 수 있다.

$$\hat{X}_n^d = [\hat{X}_{n-1}^d; \hat{X}_{\bar{n}}^e], \tag{5}$$

$n = 1,2,...,N$, $\bar{n} = N,N-1,...,1$이고, [;]는 채널 차원에서의 결합을 의미한다.

그러나 중첩 U-Net은 일반적인 U-Net과 달리 각 단계가 인코더-디코더 계층들로 구성되어 있다. 즉, 디코더 단계의 자세한 정보를 보완하기 위해서 인코더 단계의 최종 단일 출력값뿐만 아니라 디코더 단계 내의 인코더-디코더 계층에 대한 추가적인 값 전달이 필요하다. 본 논문은 이러한 정보 손실 문제를 해결하기 위해 중첩 U-Net의 구조적 특징에 적합한 MLS를 제안하였다.

그림 2는 서로 대칭되는 인코더-디코더 단계를 표현한 그림으로, 실선은 기존의 중첩 U-Net에서 사용되는 일반적인 단일 skip connection을 나타내며 점선은 본 논문에서 제안된 추가적인 skip connection을 나타낸다. 이때, 편의를 위해 다운/업 샘플링 계층과 병목 블록은 생략하였다. (a)는 인코더 단계의 디코더 계층으로부터 디코더 단계의 인코더 계층으로((d))의 실제 계층 간 연결을 나타낸 것으로 (b)~(f)는 축약해서 표시하였다. (b), (c)는 인코더 단계의 인코더 계층으로부터 각각 디코더 단계의 인코더 계층과 디코더 계층을 연결한 경우이며, (d), (e)는 인코더 단계의 디코더 계층으로부터 각각 디코더 단계의 인코더 계층과 디코더 계층을 연결한 경우이다. 그리고 (f)는 검증 데이터를 통해 (b)~(e) 간의 성능을 비교 평가하여 가장 큰 성능 향상을 보인 (d)와 (e)를 결합한 경우로 본 논문에서 제안된 MLS를 나타낸다.

디코더 단계 내의 k번째 인코더-디코더 계층의 출력이 $G^{de/dd_{n,k}}$이고 인코더 단계 내의 k번째 인코더-디코더 계층의 출력이 $G^{ee/ed_{\bar{n},k}}$이라 하면, (b), (c), (d), (e)는 각각 수식

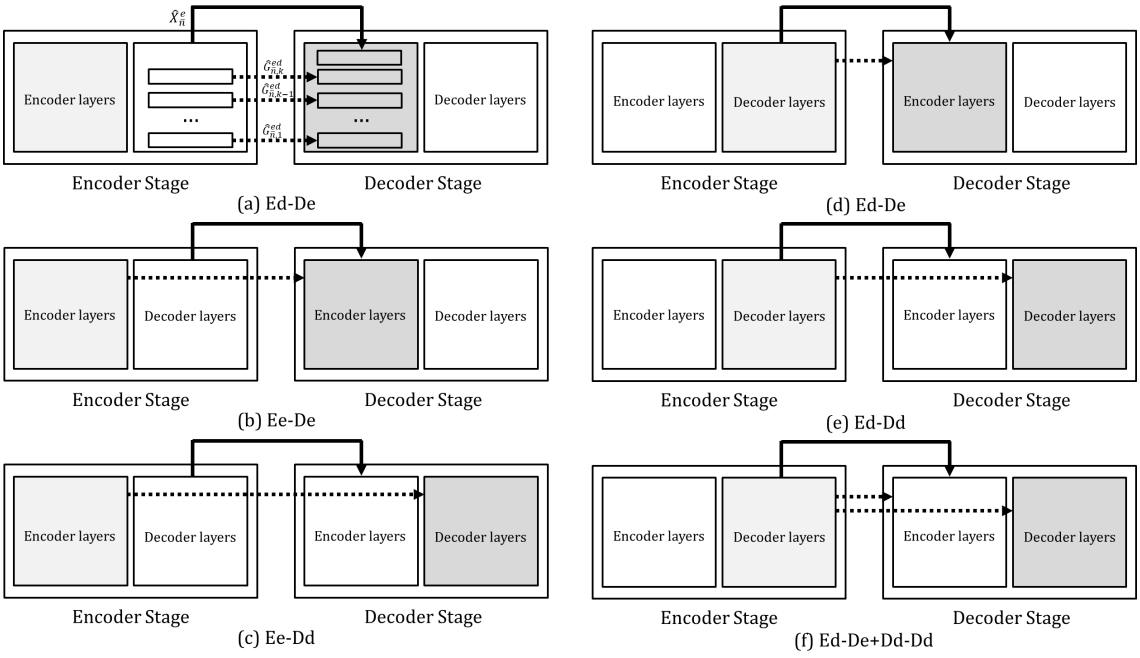


그림 2. skip connection 경로에 따른 인코더-디코더 단계. 이때, 실선은 기존의 중첩 U-Net에서 사용되는 skip connection 경로를 나타내며, 점선은 본 논문에서 제안된 추가적인 skip connection 경로를 나타낸다

Fig. 2. The encoder-decoder stages symmetrical to each other according to the skip connection path. The line indicates the standard skip connection used in the nested U-Net, and the dotted line indicates the additional skip connection proposed in this paper

(5)와 (6), (7), (8), (9)의 결합으로 표현할 수 있다.

$$G^{de_{n,k}} = [G^{de_{n,k-1}}; G^{ee^{-}_{n,i-k+1}}] \tag{6}$$

$$G^{dd_{n-1,k}} = [G^{dd_{n-1,k-1}}; G^{de^{-}_{n-1,i-k+1}}; G^{ee^{-}_{n-1,i-k+1}}], \tag{7}$$

$$G^{de_{n,k}} = [G^{de_{n,k-1}}; G^{ed^{-}_{n,i-k+}},^{1}] \tag{8}$$

$$G^{dd_{n-1,k}} = [G^{dd_{n-1,k-1}}; G^{de^{-}_{n-1,i-k+1}}; G^{ed^{-}_{n-1,i-k+1}}], \tag{9}$$

위 식에서, $k = 1,2,...,i$ 이며 i는 각 단계의 인코더-디코더 계층 수를 의미한다.

MLS는 수식 (5)와 수식(8), (9)의 결합으로 표현할 수 있다.

Ⅳ. 실험 및 결과

실험을 위한 학습용 데이터와 검증용 데이터는 깨끗한 TIMIT[10] 발화 데이터 세트에 총 11가지 종류의 잡음 데이터를 각각 Signal-to-Noise Ratio(SNR)를 0~15dB까지 1dB 단위로 섞어서 생성하였으며, 평가용 데이터는 TIMIT 발화 데이터 세트에 총 6가지 잡음 데이터를 각각 SNR 0~15 dB까지 5dB 단위로 섞어서 생성하였다. 이때, TIMIT은 630명의 영어권 화자가 각각 10개의 발화를 스튜디오 환경에서 녹음한 것이다. 학습과 검증을 위한 잡음 데이터 세트로는 핑크 노이즈, 화이트 노이즈, 배블 노이즈, 그리고 일상생활에서 발생할 수 있는 다양한 잡음으로 구성된 CHIME-2[11], CHIME-3[12], NOISEX-92[13] 데이터 세트를 사용하였다. 평가를 위한 잡음 데이터 세트로는 쇼핑센터, 라운지, 전철역 잡음 등으로 구성된 ETSI[14] 데이터 세트를 사용하였다. 모든 발화 데이터와 잡음 데이터는 16kHz로 샘플링 하였으며, 학습을 위한 데이터는 총 59,136개, 검증을 위한 데이터는 총 2,204개, 테스트를 위한 데이터는 총 1,848개를 사용하였다. 윈도우 길이, 홉 길이, FFT, 청크 길이는 각각 25ms, 6.25ms, 512 샘플, 3s를 사용하였으며, N 값으로는 6을, i 값으로는 각각의 인코더-디코더 단계별로 6, 5, 4, 4, 4, 3 값을 사용하였다. 또한, 모델 최적화를 위해 Adam optimizer와 시간-주파수 결합 손실함수[4]를 사용하였다.

제안된 음성 향상 모델의 성능을 평가하기 위한 평가 지표로는 음질을 평가하는 Perceptual Evaluation of Speech Quality(PESQ)와 각각 음성의 명료도와 배경 잡음, 그리고 전체적인 음성의 점수를 나타내는 Signal Distortion (CSIG), Background Noise Distortion(CBAK), Overall Quality (COVL)[15]를 사용하였다. 이때, PESQ는 4.5 만점, CSIG, CBAK, COVL은 5점 만점이며 높을수록 좋은 값을 의미한다.

표 1은 평가 데이터를 사용하여 일반적인 단일 skip connection을 사용하는 중첩 U-Net (Baseline)에 그림 2와 같이 추가적인 skip connection을 다양한 경로에 따라 함께 사용했을 때의 성능을 나타낸 표이다. 이때, Baseline은 참고문헌[4]의 기본 중첩 U-Net 구조의 마지막 인코더-디코더 단계의 계층 수를 조절하여 시간-주파수 영역에서 실험한 것이다.

PESQ의 경우, 인코더 단계의 인코더 계층을 각각 디코더 단계의 인코더 계층(+Ee-De)과 디코더 계층(+Ee-Dd)으로 전달했을 때 평균적으로 0.03, 0.05씩 증가한다. CSIG의 경우 평균적으로 약간 낮아지거나 같은 값을 보이지만 CBAK와 COVL는 SNR 0dB와 5dB 상황을 제외하고 같은 값을 보이거나 약간의 향상 폭을 보인다. 반면, 인코더 단계의 디코더 계층을 디코더 단계의 인코더 계층(+Ed-De)과 디코더 계층(+Ed-Dd)으로 전달했을 때는 앞선 두 경우보다 좀 더 큰 성능 향상을 보이며 모든 경우에서 Baseline보다 나은 성능을 보인다. 또한, 본 논문은 검증 데이터를 통해 가장 큰 성능 향상 폭을 보인 +Ed-De와 +Ed-Dd를 결합하여 MLS (+Ed-De+Ed-Dd)를 제안하였으며 제안된 MLS는 평가 데이터에서 가장 큰 성능 향상을 보인 +Ed-De보다 추가적인 성능 향상을 얻을 수 있다.

표 2는 MLS를 사용하는 최종 제안 모델과 최근 음성 향상에서 좋은 성적을 보인 복소 네트워크 DCCRN+C[2]와 중첩 U-Net 기반의 모델인 SADNUNet[4]과의 PESQ 성능을 평가 데이터를 사용하여 비교한 표이다. 실험 결과 제안된 모델은 DCCRN+C보다 평균적으로 0.35이상 매우 큰

표 1. skip connection의 다양한 연결 경로에 따른 성능 비교 평가표. 이때, 각각의 결괏값은 (a)PESQ, (b)CSIG, (c)CBAK, (d)COVL을 통하여 측정하였다
Table 1. Performance comparison according to various connection path of the skip connection. It measured by: (a) PESQ, (b) CSIG, (c) CBAK, and (d) COVL

(a)

Metric	Model	Param.	SNR				
			0 dB	5 dB	10 dB	15 dB	Avg.
PESQ	Noisy	-	1.20	1.41	1.73	2.18	1.63
	Baseline	2.58M	2.62	3.07	3.43	3.73	3.21
	+Ee-De	2.78M	2.62	3.08	3.46	3.80	3.24
	+Ee-Dd	2.98M	2.64	3.11	3.49	3.80	3.26
	+Ed-De	2.78M	2.73	3.19	3.55	3.85	3.33
	+Ed-Dd	2.98M	2.69	3.15	3.53	3.83	3.30
	+Ed-De+Ed-Dd	3.17M	**2.77**	**3.25**	**3.60**	**3.88**	**3.38**

(b)

Metric	Model	Param.	SNR				
			0 dB	5 dB	10 dB	15 dB	Avg.
CSIG	Noisy	-	2.92	3.33	3.70	4.14	3.52
	Baseline	2.58M	4.36	4.69	4.87	4.97	4.72
	+Ee-De	2.78M	4.30	4.67	4.87	4.98	4.71
	+Ee-Dd	2.98M	4.34	4.68	4.87	4.98	4.72
	+Ed-De	2.78M	4.40	4.73	4.90	4.99	4.76
	+Ed-Dd	2.98M	4.37	4.69	4.88	4.98	4.73
	+Ed-De+Ed-Dd	3.17M	**4.40**	**4.74**	**4.91**	**4.99**	**4.76**

(c)

Metric	Model	Param.	SNR				
			0 dB	5 dB	10 dB	15 dB	Avg.
CBAK	Noisy	-	1.55	1.95	2.40	2.94	2.21
	Baseline	2.58M	3.17	3.57	3.91	4.21	3.72
	+Ee-De	2.78M	3.20	3.61	3.98	4.30	3.77
	+Ee-Dd	2.98M	3.21	3.62	3.97	4.28	3.77
	+Ed-De	2.78M	3.26	3.66	4.00	4.30	3.81
	+Ed-Dd	2.98M	3.24	3.64	3.98	4.29	3.79
	+Ed-De+Ed-Dd	3.17M	**3.29**	**3.70**	**4.04**	**4.33**	**3.84**

(d)

Metric	Model	Param.	SNR				
			0 dB	5 dB	10 dB	15 dB	Avg.
COVL	Noisy	-	2.37	2.70	3.04	3.46	2.89
	Baseline	2.58M	3.73	4.11	4.36	4.58	4.20
	+Ee-De	2.78M	3.69	4.10	4.38	4.66	4.20
	+Ee-Dd	2.98M	3.70	4.10	4.39	4.64	4.21
	+Ed-De	2.78M	3.82	**4.20**	4.45	4.68	4.29
	+Ed-Dd	2.98M	3.70	4.12	4.41	4.65	4.22
	+Ed-De+Ed-Dd	3.17M	**3.82**	4.18	**4.46**	**4.69**	**4.29**

표 2. 제안된 모델 (+Ed-De+Ed-Dd, MLS)과 DCCRN+C, SADNUNet과의 PESQ 점수 비교 평가표
Table 2. PESQ scores of the proposed model (+Ed-De+Ed-Dd, MLS), DCCRN+C and SADNUNet

Metric	Model	Param.	SNR				
			0 dB	5 dB	10 dB	15 dB	Avg.
PESQ	Noisy	-	1.20	1.41	1.73	2.18	1.63
	DCCRN+C	3.77M	2.38	2.88	3.28	3.59	3.03
	SADNUNet	2.63M	2.54	2.97	3.31	3.60	3.11
	Proposed	3.17M	**2.77**	**3.25**	**3.60**	**3.88**	**3.38**

성능 차이를 보이며, SNR 0dB에서는 제안된 모델이 DCCRN+C보다 0.39 높은 점수를 보인다. 또한, 제안된 모델은 중첩 U-Net 기반 음성 향상 모델인 SADNUNet 보다도 평균적으로 0.27 이상 매우 높은 차이를 보이며 특히 SNR 10dB에서는 0.29의 차이를 보인다.

V. 결 론

본 논문은 중첩 U-Net 기반 음성 향상 알고리즘의 성능을 최적화하기 위해 다중 레벨 skip connection을 제안하였고 이를 위해 skip connection을 다양한 형태로 결합하여 성능을 비교 평가하였다. 제안된 MLS는 다양한 객관적 평가 지표에서 기존의 중첩 U-Net 기반 음성 향상 알고리즘 성능을 향상해주었으며 다른 심층 신경망 기반 음성 향상 모델보다도 매우 우수한 성능을 보인다.

참 고 문 헌 (References)

[1] 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*, vol. 106, p. 107404, April 2020. doi: https://doi.org/10.1016/j.patcog.2020.107404

[2] S. Zhao, T. H. Nguyen, and B. Ma, "Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses," in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6648 – 6652, 2021.

[3] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6633 – 6637, 2021.

[4] X. Xiang, X. Zhang, and H. Chen, "A nested u-net with self-attention and dense connectivity for monaural speech enhancement," *IEEE Signal Processing Letters*, vol. 29, pp. 105 – 109, 2022. doi: https://doi.org/10.1109/LSP.2021.3128374

[5] S.-R. Hwang, S.-W. Park, and Y.-C. Park, "Performance comparison evaluation of real and complex networks for deep neural network-based speech enhancement in the frequency domain." *The Journal of the Acoustical Society of Korea*, vol. 41, no. 1, pp. 30-37, 2022. doi: https://doi.org/10.7776/ASK.2022.41.1.030

[6] Y. Xian, Y. Sun, W. Wang, and S. M. Naqvi, "A multi-scale feature recalibration network for end-to-end single channel speech enhancement," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 143 – 155, 2021. doi: https://doi.org/10.1109/JSTSP.2020.3045846

[7] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation." in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1055 – 1059, 2020.

[8] 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*. 2019. doi: https://doi.org/10.48550/arXiv.1903.03107

[9] H. Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708, 2017.

[10] J. W. Lyons, DARPA TIMIT acoustic-phonetic continuous speech corpus, 1993.

[11] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'chime'speech separation and recognition challenge: Datasets, tasks and baselines," in IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 126 – 130, 2013.

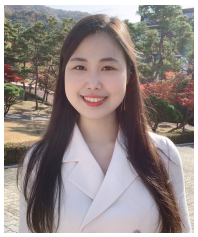
[12] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, taskand baselines," in IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, pp. 504 – 511, 2015.

[13] 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., vol. 12, no. 3, p. 247 – 251, 1993. doi: https://doi.org/10.1016/0167-6393(93)90095-3

[14] ETSI, 202 396-1: Speech quality performance in the presence of background noise, 2009.

[15] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229 – 238, 2008. doi: https://doi.org/10.1109/TASL.2007.911054

저 자 소 개



황 서 림

- 2017년 : 연세대학교 컴퓨터정보통신공학부
- 2021년 : 연세대학교 일반대학원 전산학과 석박사통합과정
- ORCID : http://orcid.org/0000-0002-7407-2440
- 주관심분야 : 음성향상, 음성신호처리, 디지털신호처리, 딥러닝

변 준

- 2017년 : 연세대학교 컴퓨터정보통신공학부
- 2021년 : 연세대학교 일반대학원 전산학과 석박사통합과정
- ORCID : http://orcid.org/0000-0002-5114-5459
- 주관심분야 : 오디오코딩, 오디오신호처리, 음성신호처리, 딥러닝

허 준 영

- 2016년 : 강릉원주대학교 전자공학과
- 2022년 : 연세대학교 일반대학원 전산학과 석사과정
- ORCID : http://orcid.org/0000-0003-0665-784X
- 주관심분야 : 능동소음제어, 오디오신호처리, 음성신호처리, 딥러닝

차 재 빈

- 2017년 : 연세대학교 컴퓨터정보통신공학부
- ORCID : http://orcid.org/0000-0002-9294-8362
- 주관심분야 : 음성향상, 음성신호처리, 디지털신호처리, 딥러닝

박 영 철

- 1982년 : 연세대학교 전자공학과 학사
- 1986년 : 연세대학교 전자공학과 석사
- 1988년 : 연세대학교 전자공학과 박사
- 2002년 : 연세대학교 소프트웨어학부 교수
- ORCID : http://orcid.org/0000-0003-3274-076X
- 주관심분야 : 디지털신호처리, 음성신호처리, 적응신호처리, 오디오신호처리, 딥러닝