Nested U-Net 기반 잡음 제거를 위한 two-level skip connection 제안 및 성능 비교 평가

*황서림 *변준 *허준영 *차재빈 *박영철†
*연세대학교 지능형신호처리연구실
young00@yonsei.ac.kr†

Performance comparative evaluation of Two-level skip connection for nested U-Net-based noise cancellation.

Hwang, Seorim Byun, Joon Heo, Junyeong Cha, Jaebin Park, Youngcheol
Intelligent Signal Processing Lab, Yonsei University

요약

본 논문은 최근 잡음 제거에서 우수한 성능을 보인 Nested U-Net의 성능을 최적화하기 위하여 두 단계로 이루어진 two-level skip connection (TLS)을 제안하였다. 이때, 인코더와 디코더의 경로를 다르게 하여 다양한 형태의 TLS을 제안하고 각 형태의 성능을 비교 평가하였다. 또한, 가장 좋은 성능을 보인 두 개의 경로를 조합하여 최종 Nested U-Net 기반 모델을 제안하였다. 제안된 모델은 다른 잡음 제거 모델과 비교하여 객관적인 평가 지표에서 매우 우수한 성능을 보인다.

1. 서론

잡음 제거는 불필요한 배경 잡음을 제거하여 깨끗한 음성을 복원해 내는 기술로 음성 인식 인공지능과 보청기 등 음질과 음성의 명료도가 중요한 응용 분야에 필수적으로 사용되고 있다. 최근에는 딥 러닝의 발전과 함께 딥 러닝 기반 잡음 제거 기술이 많이 연구되고 있으며, 기존의 확률 기반 기법과 비교하여 매우 우수한 성능을 보이고 있다[1-4].

딥 러닝 기반 잡음 제거에서 음성의 문맥 정보를 고려하여는 것은 모델의 성능과 밀접한 연관성을 가진다. 최근에는 multi-scale을 사용하여 음성의 문맥 정보를 고려하는 시도가 늘고 있으며[3][4], Nested U-Net은 이러한 multi-scale을 성공적으로 사용한 네트워크이다[1][4]. 그러나 현재까지 제안된 Nested U-Net은 중첩된 네트워크 구조에 일반적인 U-Net에서 사용되는 단일 skip connection을 사용하고 있다.

본 논문은 Nested U-Net의 성능을 최대화하기 위하여 중첩된 네트워크 구조에 적합한 two-level skip connection (TLS)을 제안하였다. 이때, TLS는 인코더와 디코더 경로에 따라 다양한 형태로 연결하여 성능을 비교 평가하였으며 제안된 TLS를 사용한 모델은 기존의 skip connection과 다른 state-of-the-art (SOTA) 모델과 비교하여 우수한 성능을 보인다.

2. Nested U-Net과 two-level skip connection

Nested U-Net은 U-Net의 각 계층을 U 모양의 잔차 블록으로 변경한 네트워크[1][4]로 구조는 Figure. 1과 같다. Nested U-Net은 잡음이 섞인 시간 영역의 음성 y가 입력으로 들어오면 Short-Time Fourier Transform (STFT) 을 통해 시간-주파수 영역으로(Y) 변환한다. 그리고 Y의 크기 성분 $|Y|$를 각각의 인코더 단계와 바틀넥 블록, 그리고 디코더 단계를 통해 잡음이 제거된 음성의 크기 성분 $|\hat{X}|$를 구한다. $\hat{X}$는 모델의 출력 $|\hat{X}|$와 입력으로 들어온 Y의 위상 정보(θ_Y)를 결합하여 구해지며 Inverse STFT (ISTFT)를 통해 최종적으로 시간 영역의 향상된 음성($\hat{x}$)을 얻는다.

기존의 Nested U-Net은 일반적인 U-Net과 동일하게 디코더 단계의 업 샘플링 과정에서 디테일한 정보를 보완하기 위하여 인코더 단계로부터 값을 직접적으로 전달 받는다 (top-level skip connection) [1][4]. 그러나 이러한 단일 skip connection은 Nested U-Net의 각 계층 내의 인코더-디코더의 디테일한 정보를 보완하지 못한다. 본 논문은 이러한 Nested U-Net의 구조적 가능성을 활용하기 위하여 Figure. 2와 같이 큰 U-Net 안의 U 모양의 잔차 블록 간의 추가적인 skip connection을 사용하였으며 다양한 경로에 따른 two-level skip connection을 제안하여 실험 및 비교 분석하였다.

3. 실험 결과 및 분석

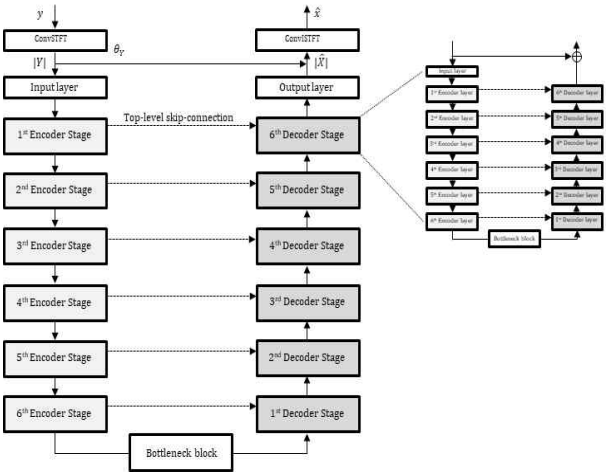

Figure 1. The architecture of the Nested U-Net.

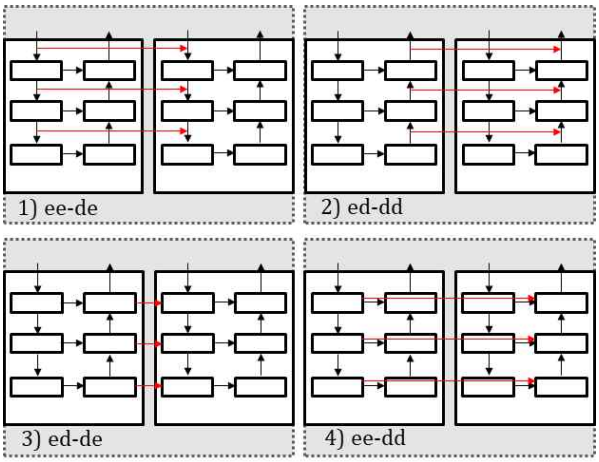

Figure 2. Skip connections along the forwarding path. 1) From the encoder path of the encoder stage to the encoder path of the decoder stage, 2) From the decoder path of the encoder stage to the decoder path of the decoder stage, 3) From the encoder path of the decoder stage to the encoder path of the decoder stage , and 4) from the decoder path of the decoder stage to the encoder path of the decoder stage.

실험을 위한 데이터로는 깨끗한 TIMIT 발화 데이터 셋에 총 11가지 종류의 잡음 데이터를 각각 랜덤하게 뽑아 Signal-to-Noise Ratio (SNR) 0~15 dB 까지 1 dB 단위로 생성하였다. 이때, 잡음 데이터 셋은 훈련을 위해 CHIME-2, CHIME-3, NOISEX-92 데이터 셋을 사용하였으며, 테스트를 위해 ETSI 데이터 셋을 사용하였다. 모든 발화는 16k로 샘플링 하였으며, 훈련을 위한 데이터는 총 59,136개, 테스트를 위한 데이터는 총 2,304개를 사용하였다. 윈도우 길이, 홉 길이, FFT는 각각 25 ms, 6.25 ms, 512 샘플을 사용하였으며, 모델 최적화를 위해서 Adam optimizer와 시간-주파수 결합 손실함수를 사용하였다.

잡음 제거 모델의 성능 평가를 위한 객관적 평가 지표로는 perceptual evaluation of speech quality (PESQ)을 사용하였다. 총 네 가지 형태의 TLS를 베이스라인 네트워크에 추가하여 학습한 모델과 가장 성능이 좋은 두 형태를 결합하여 학습한 모델의 성능을 비교 평가하였다. 그리고 가장 좋은 성능을 보인 모델을 다른 딥 러닝 기반 잡음 제거 모델들[2-4]과 비교하였으며 결과는 각각 Table. 1과 Table 2와 같다.

실험 결과 단일 skip connection을 사용한 것보다 TLS를 사용한 경우 모든 SNR 상황에서 성능 향상이 있었다. 특히, 큰 U-Net의 인코더 단계의 디코더 경로를 디코더 단계로 연결한 경우 (+ ed-de, + ed-dd) 베이스라인과 비교하여 평균적으로 PESQ 0.09이상 향상되었다. 최종적으로 두 경로를 함께 사용하여 제안된 구조는 (+ ed-de + ed-dd, Proposed) 베이스라인과 비교하여 평균적으로 PESQ가 0.17 이상 증가하는 것을 확인하였다. 그리고 이러한 제안된 구조는 다른 딥 러닝 기반 잡음 제거 모델과 비교하여서도 우수한 성능을 보였다.

Table 1. Performance evaluation for ablation test.

Model (param.)	PESQ			
	0 dB	5 dB	10 dB	15 dB
Noisy	1.20	1.41	1.73	2.18
Baseline (2.58M)	2.62	3.07	3.43	3.73
+ Ee-De (2.78M)	2.62	3.08	3.46	3.80
+ Ed-Dd (2.98M)	2.69	3.15	3.53	3.83
+ Ed-De (2.78M)	2.73	3.19	3.55	3.85
+ Ee-Dd (2.98M)	2.64	3.11	3.49	3.80
+ Ed-De + Ed-Dd (3.17M)	**2.77**	**3.25**	**3.60**	**3.88**

Table 2. Performance evaluation with other state-of-the-art speech denoising model.

Model (param.)	PESQ			
	0 dB	5 dB	10 dB	15 dB
Noisy	1.20	1.41	1.73	2.18
DCCRN+C [2] (3.77M)	2.38	2.88	3.28	3.59
FullSubNet [3] (5.64M)	2.28	2.71	3.16	3.49
SADNUNet [4] (2.63M)	2.54	2.97	3.31	3.60
Proposed (3.17M)	**2.77**	**3.25**	**3.60**	**3.88**

4. 결론

본 논문은 다양하게 조합한 two-level skip connection을 통해 음성의 문맥 정보를 활용하는데 유용한 Nested U-Net의 성능을 최적화하였다. 실험 결과 큰 U-Net의 인코더 단계의 디코더 경로를 디코더 단계로 전해주는 것이 가장 큰 성능 향상을 보였으며, 두 경로를 함께 사용하

였을 때 추가적인 성능 향상을 얻었다. 제안된 구조를 사용한 모델은 다른 딥 러닝 기반 잡음 제거 모델과 비교하여 평균적으로 PESQ 0.27이상 매우 우수한 성능을 보인다.

참고문헌

[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, 2020.

[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), 2021, pp. 6648-6652.

[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), 2021, pp. 6633-6637.

[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.