한국음향학회지 제42권 제2호 pp. 94~101 (2023)
The Journal of the Acoustical Society of Korea Vol.42, No.2 (2023)
https://doi.org/10.7776/ASK.2023.42.2.094

pISSN : 1225-4428
eISSN : 2287-3775

복소 스펙트럼 기반 음성 향상의 성능 향상을 위한 time-frequency self-attention 기반 skip-connection 기법 연구

A study on skip-connection with time-frequency self-attention for improving speech enhancement based on complex-valued spectrum

정재희,[1] 김우일[1†]

(Jaehee Jung[1] and Wooil Kim[1†])

[1]인천대학교 컴퓨터공학부

(Received January 25, 2023; revised February 27, 2023; accepted March 8, 2023)

초 록: 음성 향상에서 많이 사용되는 U-Net과 같이 인코더와 디코더로 구성된 심층 신경망 모델은 skip-connection을 통해 인코더의 특징을 디코더에 연결하는 구조로 구성되어 있다. Skip-connection은 디코더에서 향상된 스펙트럼을 재구성하는데 도움을 주며 인코더를 통해 손실된 정보를 보완해줄 수 있다. 이때 skip-connection을 통해 연결되는 인코더의 특징과 디코더의 특징의 의미는 서로 다르다. 본 논문에서는 복소 스펙트럼 기반 음성 향상의 성능 향상을 위해 디코더에 연결되는 인코더의 특징을 디코더 특징의 의미에 가깝게 변환해주도록 skip-connection에 Self-Attention(SA)을 적용하는 방안을 연구하였다. SA는 시퀀스-시퀀스 문제에서 출력 시퀀스를 생성할 때, 입력 시퀀스의 가중 산술 평균을 이용하여 결정적인 부분을 집중해서 볼 수 있도록 하는 기법으로, 음성 향상 분야에서도 이를 적용함으로써 성능 향상에 효과적임을 입증하는 연구가 진행되었다. SA를 skip-connection에 적용하기 위해 인코더 특징과 디코더 특징을 이용하는 총 3가지의 방법에 대해 연구하였다. TIMIT 데이터베이스를 이용한 음성 향상 실험 결과, 제안하는 방법이 기존 skip-connection으로만 연결된 Deep Complex U-Net(DCUNET)과 비교하여 모든 성능 평가 지표에서 향상된 결과를 보였다.

핵심용어: 복소 스펙트럼, 음성 향상, Skip-connection, Self-Attention

ABSTRACT: A deep neural network composed of encoders and decoders, such as U-Net, used for speech enhancement, concatenates the encoder to the decoder through skip-connection. Skip-connection helps reconstruct the enhanced spectrum and complement the lost information. The features of the encoder and the decoder connected by the skip-connection are incompatible with each other. In this paper, for complex-valued spectrum based speech enhancement, Self-Attention (SA) method is applied to skip-connection to transform the feature of encoder to be compatible with the features of decoder. SA is a technique in which when generating an output sequence in a sequence-to-sequence tasks the weighted average of input is used to put attention on subsets of input, showing that noise can be effectively eliminated by being applied in speech enhancement. The three models using encoder and decoder features to apply SA to skip-connection are studied. As experimental results using TIMIT database, the proposed methods show improvements in all evaluation metrics compared to the Deep Complex U-Net (DCUNET) with skip-connection only.

Keywords: Complex-valued spectrum, Speech enhancement, Skip-connection, Self-Attention

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)

I. 서 론

잡음에 오염된 음성을 음성 인터페이스에 효과적으로 이용하기 위해 음성 향상 기술을 적용할 수 있다. 음성 향상은 잡음 오염 음성에서 잡음을 최대한 억제해 음성의 지각적인 품질과 명료도를 최대화하도록 수행한다.

음성 향상은 통계적인 기법 기반 기술로 시작하여 심층 신경망 기반의 기술로 연구되어왔다. 통계적인 기법 기반 기술로서 스펙트럼 차감법, 위너 필터 등이 연구되었다.[1,2] 통계적인 기법 기반의 음성 향상은 시간에 따라 통계 정보가 변하는 비정상 잡음 오염 음성이나 낮은 Signal-to-Noise Ratio(SNR)의 잡음 오염 음성에 대해서는 성능이 많이 떨어진다. 이를 해결하기 위해 심층 신경망을 이용한 음성 향상 기법이 연구되었다.[3-5]

심층 신경망 기반의 음성 향상은 심층 신경망 모델의 입력으로 잡음 오염 음성을 이용하여 향상된 음성 또는 마스크를 추정한다. 이때 사용하는 음성 특징은 시간 영역의 파형 또는 주파수 영역의 스펙트럼을 이용한다. 가장 많이 연구된 방법은 크기 스펙트럼을 이용한 마스크 기반의 음성 향상 시스템으로 모델을 통해 추정된 마스크는 잡음 오염 음성의 크기 스펙트럼에 곱해 향상된 음성의 크기 스펙트럼을 얻을 수 있고, 잡음 오염 음성의 위상을 그대로 사용해 향상된 음성 파형을 얻을 수 있다. 최근, 음성 향상 학습에서 크기 스펙트럼뿐만 아니라, 위상의 중요성이 강조되면서 복소 스펙트럼을 이용하여 음성 향상을 수행하는 방안에 대해 많이 연구되고 있다.[6,7]

최근에는 시퀀스-투-시퀀스 문제에서 출력 시퀀스 생성을 위한 학습을 효과적으로 수행할 수 있도록 attention을 적용하는 방법이 연구되어 많이 활용되었다.[8-10] 대표적으로는 Self-Attention(SA) 방법이 많이 사용되고 있다. SA는 입력 시퀀스를 이용해 출력 시퀀스를 생성할 때 입력 시퀀스의 가중 산술 평균을 이용하여 결정적인 부분을 더욱 주의집중하여 볼 수 있도록 하는 기법이다. 음성 향상 분야에서도 스펙트럼의 상관관계가 잡음 제거 성능에 중요하다는 것이 발견되어 SA가 많이 활용되었다.[10,11] 최근, 스펙트럼의 시간 축, 주파수 축에 따라 스펙트럼을 재정렬하여 병렬로 수행하는 SA 기법이 제안되었다.[10,12] TFT-Net의 성능 향상을 위해 시간 및 주파수 축을 따라 병렬로 상관성을 계산하는 Sample-Independent Dual Attention Block(SDAB)[12] 기법이 제안되었고, 이를 기반으로 하여 주파수 영역에서 스펙트럼 요소 사이의 상관관계를 학습할 때 발생하는 높은 복잡성을 해결하기 위해 Time-Frequency SA(TFSA) 기법이 제안되었다.[10]

현재 음성 향상에서 많이 사용되는 U-Net과 같이 인코더와 디코더로 구성된 심층 신경망 구조에서 사용되는 skip-connection은 인코더의 특징을 디코더에 연결하여 사용되며, 인코더에 의해 특징이 압축되면서 손실된 정보를 보완하고 디코더에서 향상된 스펙트럼을 재구성하는데 도움을 준다. 그러나 skip-connection은 잡음 오염 음성 스펙트럼에 가까운 첫 인코더의 특징을 최종 향상된 스펙트럼을 재구성하는 디코더에 연결한다. 이와 같이 인코더의 특징과 디코더의 특징은 엄연히 서로 다른 특징으로 볼 수 있다. 이를 보완하기 위해 Skip Convolutional Neural Network(SkipConvNet)이 제안되었다.[13] SkipConvNet은 U-Net의 skip-connection 중간에 컨볼루션 층을 추가하여 인코더 특징을 변환하여 디코더에 연결한다.

본 논문에서는 복소 스펙트럼 기반 음성 향상 모델의 성능 향상을 위해서 U-Net 구조와 같은 심층 신경망 구조의 skip-connection에서 발생하는 인코더와 디코더 특징의 불일치를 줄이도록 SA를 적용하는 방안에 관한 연구를 수행하였다. SA는 시간 축, 주파수 축에 따라 병렬로 수행하도록 TFSA를 이용하였고 복소 스펙트럼 기반 음성 향상을 수행하기 위해 Deep Complex U-Net(DCUNET)[4] 모델을 이용하였다. Skip-connection에 TFSA를 적용하기 위해 인코더의 특징만 이용하는 방법과 디코더의 특징과 가깝도록 변환하기 위해 인코더 특징과 디코더 특징을 같이 이용하는 방법에 대해 연구를 수행하였다.

다음 2장에서는 복소 스펙트럼 기반 음성 향상에 대한 잡음 오염 생성 및 수행 과정과 DCUNET 모델 구조에 대해 설명하고, 3장에서는 TFSA 모델의 구조를 설명한다. 4장에서는 기존 SkipConv에 대한 설명과 skip-connection에 TFSA를 적용할 수 있도록 제안하는 방법에 대해 설명한다. 5장에서는 실험과 그 결

과에 대해 논의하고 6장에서 결론을 맺는다.

II. 복소 스펙트럼 기반 음성 향상

2.1 잡음 오염 음성 생성

잡음에 오염되는 음성은 다음과 같이 시간 영역에서 깨끗한 음성 파형과 잡음 파형을 합하여 생성되는 것으로 가정한다.

$$x = s + d, \tag{1}$$

여기서 x는 생성된 잡음 오염 음성으로 s는 깨끗한 음성, d는 배경 잡음이다. 잡음 오염 음성 파형은 주파수 영역의 스펙트럼으로 변환하여 다음과 같은 식으로 표현할 수 있다.

$$X = S + D. \tag{2}$$

$$X = X_r + jX_i. \tag{3}$$

Eq. (2)에서 X, S, D는 각각 잡음 오염 음성과 깨끗한 음성, 배경 잡음의 복소 스펙트럼으로 Eq. (3)과 같이 실수부 r과 허수부 i로 구성되어 있다.

2.2 복소 스펙트럼 기반 음성 향상 과정

본 논문에서 수행한 복소 스펙트럼을 이용한 마스크 기반 음성 향상은 Fig. 1과 같이 수행된다. 시간 영역의 잡음 오염 음성 파형은 단시간 푸리에 변환을 통해 복소 스펙트럼으로 변환한 뒤 심층 신경망 모델의 입력으로 사용된다. 심층 신경망 모델은 마스크를 추정하며, 이를 잡음 오염 음성의 복소 스펙트럼과 곱하여 향상된 복소 스펙트럼을 얻을 수 있다. 향상된 복소 스펙트럼은 역 단시간 푸리에 변환을 통해 시간 영역의 향상된 음성으로 변환할 수 있다.

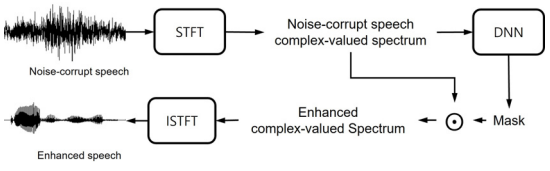

Fig. 1. Process of mask-based speech enhancement using complex-valued spectrum.

향상된 복소 스펙트럼은 심층 신경망 모델을 통해 추정된 마스크와 잡음 오염 음성의 복소 스펙트럼과의 요소 곱 연산을 통해 얻을 수 있는데, 다음과 같이 복소 스펙트럼을 크기와 위상 성분으로 변환하여 계산하였다.

$$\hat{S} = |X| \cdot |M| \cdot e^{j(X_\theta + M_\theta)}. \tag{4}$$

Eq. (4)에서 $\hat{S}$는 향상된 복소 스펙트럼이고, M는 추정된 마스크로 Eq. (3)과 같은 복소 형태를 가진다. $|\cdot|$는 크기 스펙트럼을 나타내고, θ는 위상 스펙트럼을 나타낸다. $|X|$와 $|M|$은 각각 잡음 오염 음성의 크기 스펙트럼과 추정된 마스크의 크기 성분이고, X_θ와 M_θ은 각각 잡음 오염 음성과 마스크의 위상을 나타낸다.

2.3 음성 향상 모델 구조

복소 스펙트럼 기반 음성 향상을 위해 DCUNET 모델[4]을 사용하였다. DCUNET은 Fig. 2와 같은 구조로 구성되어 있다.

인코더와 디코더는 각각 8계층으로, 각 인코더는 컨볼루션과 배치 정규화, 활성화 함수로 구성되어 있다. 디코더는 인코더와 비슷한 구조로 컨볼루션 대신 전치 컨볼루션을 수행한다. DCUNET 모델은 인코더를 통해 입력 특징을 압축하고, 디코더를 통해 이를 재구성한다. 인코더와 디코더 사이에는 skip-connection으로 연결되어있어, 이를 통해 기울기 소멸 문제를 방지하고 원 스펙트럼의 정보를 전송할

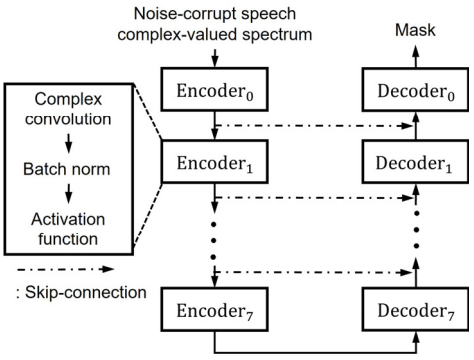

Fig. 2. A structure of DCUNET model.

수 있다.[14,15]

복소 행렬을 이용해 심층 신경망을 학습하는 경우 계산 비용이 많이 들기 때문에 복소 스펙트럼의 실수부와 허수부를 이용해서 심층 신경망 연산을 복소 연산으로 변환하여 학습을 수행할 수 있다.

DCUNET 모델은 기존 컨볼루션 대신 복소 컨볼루션을 수행한다. 컨볼루션 필터를 복소 형태를 가지는 W라고 가정했을 때, 복소 컨볼루션은 다음과 같이 연산을 수행한다.[4,5]

$$W = W_r + jW_i. \tag{5}$$

$$Y = (W_r * X_r - W_i * X_i) + j(W_r * X_i + W_i * X_r). \tag{6}$$

Eq. (6)에서 *은 컨볼루션 연산이다.

DCUNET 모델에서 사용한 배치 정규화 및 활성화 함수는 Reference [4] 논문과 동일하게 구현하여 음성 향상을 수행하였다. 사용한 활성화 함수는 마지막 디코더 층을 제외하고 Leaky-Rectified Linear Unit(ReLU)를 사용하였으며, 마지막 디코더 층에서는 마스크 값의 범위를 지정하기 위해 Tanh 활성화 함수를 이용하였다.

III. Time-frequency self attention

TFT-Net[12]은 주파수 영역과 시간 영역의 단점을 보완하기 위해 교차 영역에서 음성 향상을 수행하는 네트워크로 잡음 제거 성능에 중요한 시간-주파수 스펙트럼의 장거리 상관관계를 고려하기 위해 SDAB를 설계하였다. SDAB는 스펙트럼의 장거리 상관관계를 고려할 때 발생하는 높은 복잡성을 해결하기 위해 시간 및 주파수 축에 따른 상관관계를 병렬로 활용한다. 음성 신호는 시계열 데이터이기 때문에 시간 축을 따라 전체적으로 상관관계가 존재하고 주파수 축을 따라 고조파 상관관계가 존재한다. 이와 같이 시간 축, 주파수 축으로 상관관계를 고려하면 높은 복잡성을 해결하면서 음성 향상 성능을 향상시킬 수 있다.

이를 기반으로 Reference [10] 논문에서는 SDAB에서 사용된 완전-연결 계층을 대신하여 현재 음성 향상 분야에서 많이 사용되는 SA를 시간 축, 주파수 축에 따라 재정렬하여 수행하였다.

SA은 입력 시퀀스의 전후관계를 고려하여 얻은 가중치를 이용해 출력 시퀀스를 생성하는 기법으로 학습된 가중치는 입력 시퀀스의 내용 중에서 집중해야 하는 수준을 나타낸다.[11]

Reference [10] 논문의 TFSA는 Fig. 3과 같은 구조를 갖는다. 우선, 입력 스펙트럼을 시간 축 또는 주파수 축으로 재정렬한 뒤에 SA를 수행한다. SA는 아래의 식과 같이 입력 스펙트럼의 선형 변환된 값인 쿼리, 키, 값을 이용해 계산된다. 쿼리와 키는 내적을 계산한 후 Softmax 함수를 이용해 attention 가중치를 계산하고, 얻은 가중치와 값을 내적하여 최종 가중된 결과를 얻을 수 있다.

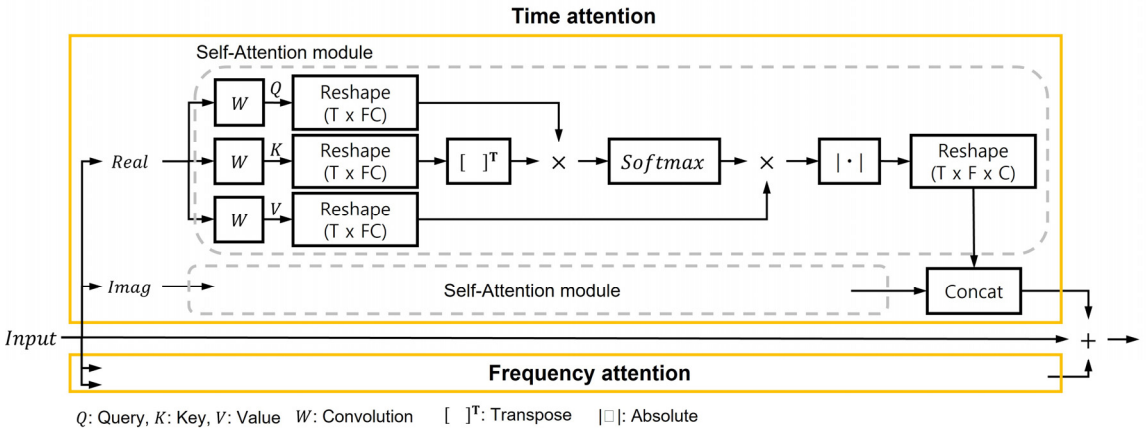

Fig. 3. (Color available online) A structure of TFSA.

$$Corr = QK^{\mathbf{T}}, \quad A = \frac{\exp^{Corr(i,j)}}{\sum_{j} \exp^{Corr(i,j)}}, \quad O = AV. \quad (7)$$

사용되는 스펙트럼은 복소 형태로 TFSA를 적용할 때, 복소 값의 실수부, 허수부 각각에 적용한다.

IV. Skip-connection attention

4.1 SkipConvNet[13]

U-Net 모델은 skip-connection을 통해 인코더 특징을 디코더에 연결해줌으로써 압축된 특징으로 인해 손실된 정보를 보완하고 디코더에서 향상된 스펙트럼을 재구성하는데 도움을 준다.

그러나 skip-connection은 잡음 오염 음성과 가까운 처음의 인코더 특징이 최종 출력과 가까운 디코더에 연결된다. 이렇게 서로 다른 인코더와 디코더의 특징을 그대로 연결하여 사용하는 경우 학습 능력이 제한될 수 있다. SkipConvNet에서는 디코더 학습에 직접적으로 도움이 될 수 있도록 skip-connection을 통해 연결되는 인코더 특징을 변환하는 새로운 컨볼루션 층을 제안하였다. SkipConvNet 모델의 skip-connection에 적용된 SkipConv 구조는 컨볼루션을 수행하고 잔차 연결로 구성되어 있으며 이를 DCUNET 모델에 적용한 구조를 Fig. 4에서 볼 수 있다.

4.2 Skip-TFSA

SkipConvNet과 동일하게 skip-connection에서 발생할 수 있는 인코더 특징과 디코더 특징 사이에 발생하는 불일치 문제를 해결하여 음성 향상의 성능을 향상시키기 위해 skip-connection에 SA를 적용하는 방안에 대해 연구를 수행하였다. SA는 시퀀스-시퀀스 문제에서 출력 시퀀스를 생성할 때, 입력 시퀀스에서의 결정적인 부분을 집중적으로 볼 수 있도록 도와주기 때문에 이를 skip-connection에 적용하여 디코더에서 향상된 스펙트럼을 재구성할 때 더욱 도움을 줄 수 있다. 또한 SA를 시간 축에 따른 상관관계와 주파수 축의 상관관계를 고려하기 위해 TFSA와 동일하게 시간 축, 주파수 축으로 재정렬하여 SA를 적용하였다.

TFSA를 skip-connection에 적용하기 위해 총 3개의 모델에 대한 연구를 수행하였다. 제안하는 3가지 모델에서 사용되는 attention은 모두 Fig. 3의 TFSA의 구조를 이용하였다.

첫 번째는 skip-connection에서 사용되는 인코더 특징만 이용하여 TFSA를 적용하는 방법으로 Fig. 5에서 구조를 볼 수 있다. Skip-connection에 연결되는 인코더 특징을 TFSA의 입력으로 사용하여 1x1 컨볼루션 연산을 통해 쿼리, 키, 값으로 변환하여 SA 연산을 수행한다.

두 번째 방법은 인코더 특징뿐만 아니라, 디코더의 특징을 같이 이용하는 방법으로 skip-connection에서 발생하는 인코더 특징과 디코더 특징의 불일치를 해결하기 위해 인코더 특징뿐만 아니라 디코더의 특징을 같이 사용하고자 한다. 제안하는 방법의 전체적인 구조는 Fig. 6에서 볼 수 있다. 해당 방법은 TFSA에서 사용되는 쿼리, 키, 값 중에서 attention 가중치를 계산하는데 사용되는 쿼리를 디코더 특징으로 사용하고, 키와 값은 인코더의 특징을 사용한다. 디코더 특징을 쿼리로 사용하고 키는 인코더 특징을 사용함으로써 attention 가중치를 계산할 때, 인코더 특징과 디코더 특징의 상관관계를 고려할 수 있다.

마지막으로 제안하는 방법은 인코더와 디코더 특

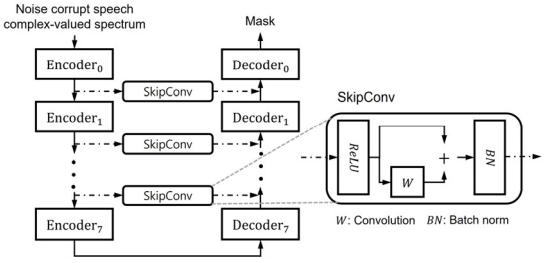

Fig. 4. A structure of DCUNET applied SkipConv.

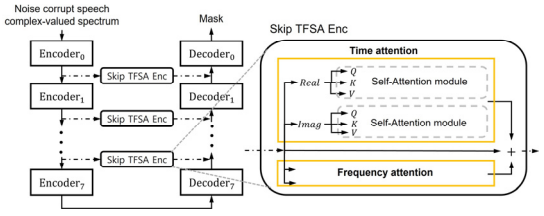

Fig. 5. (Color available online) A structure of DCUNET applied TFSA. The input of TFSA uses only encoder features (Skip TFSA Enc).

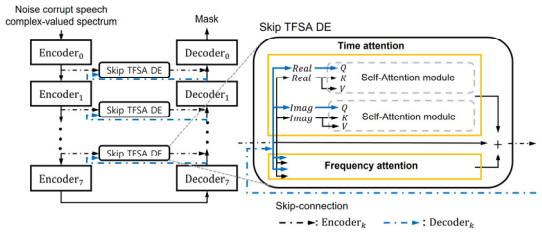

Fig. 6. (Color available online) A structure of DCUNET applied TFSA. The key and value in TFSA use encoder features, and the query uses decoder features (Skip TFSA DE).

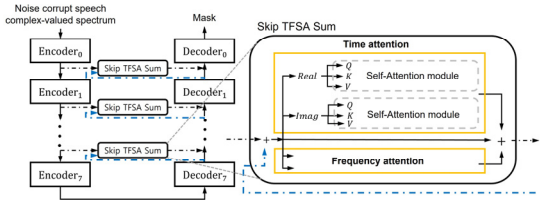

Fig. 7. (Color available online) A structure of DCUNET applied TFSA. The input of TFSA uses the sum of encoder and decoder features (Skip TFSA Sum).

징을 더한 후에 TFSA를 적용하는 방안으로 인코더 특징과 디코더 특징을 통합해주기 위해 두 특징을 더해준 후에 TFSA를 적용하였다. 해당 방법을 통해 인코더 특징과 디코더 특징이 함께 고려되어 두 특징의 불일치를 줄일 수 있음을 기대할 수 있다. Skip-connection에 적용한 모델의 전체적인 구조는 Fig. 7에서 볼 수 있다.

4.3 손실함수

모델 학습을 위해 Scale Invariant-Source to Noise Ratio(SI-SNR) 손실함수를 이용하였다.[16]

SI-SNR 손실함수는 향상된 복소 스펙트럼을 역 단시간 푸리에 변환을 통해 향상된 음성 파형으로 변환한 뒤에 계산할 수 있다. 향상된 음성 파형에서 깨끗한 음성 부분을 제거하여 얻은 잡음과 크기가 조정된 깨끗한 음성과의 비율을 계산해 손실함수 값을 얻을 수 있다.

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

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

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

Eq. (8)에서 $<\cdot,\cdot>$은 두 벡터에 대한 내적 곱 연산이고, $\|\cdot\|_2$는 L2 정규화 연산이다. $\hat{s}$은 향상된 음성 파형이고, s는 깨끗한 음성 파형이다.

V. 실험 및 결과

5.1 데이터베이스

잡음 오염 음성 생성을 위한 깨끗한 음성 데이터는 TIMIT 데이터베이스[17]를 사용하였고 배경 잡음은 한국 TV 프로그램들을 사용하였다. 사용한 배경 잡음의 종류는 '드라마', '뉴스', '음악', '스포츠'로 가정에서 발생할 수 있는 종류의 잡음을 사용하였다. 잡음 오염 음성은 배경 잡음에 Room Impulse Response (RIR) 필터를 적용한 후 깨끗한 음성과 SNR 5 dB, 0 dB, -5 dB 조건으로 합하여 생성하였다. 사용한 RIR 필터는 마이크와의 거리 1 m, 2.5 m, 4 m인 경우와 잔향 시간이 0.3 s, 0.4 s, 0.5 s일 때의 조합으로 구성하였다. 생성된 잡음 오염 음성 데이터 중 훈련 데이터는 55,440 발화, 검증 데이터는 1,200 발화, 테스트 데이터는 2,304 발화를 사용하였다.

사용한 모든 음성은 16 KHz 이며, 윈도우 크기 및 이동 크기는 32 ms와 16 ms로 설정하였다. 푸리에 변환은 512로, 주파수 256차원과 에너지 값을 포함하여 총 257차원으로 사용하였다.

5.2 실험 설정

모델의 학습을 위해 'Adam' 최적화 알고리즘을 사용하여 학습률은 0.001로 설정하였다. 또한 배치는 16으로 설정하여 학습을 수행하였다.

제안하는 방안의 성능 평가 비교를 위해 UNET 모델과 DCUNET 모델을 평가하였다. UNET 모델은 DCUNET 모델과 동일한 구조의 모델로, 복소 스펙트럼 대신 크기 스펙트럼을 이용하여 음성 향상을 수행하였다.

모델의 성능 평가를 위해서 총 3가지의 성능 평가 지수를 사용하였다. 첫 번째는 Source-to-Distortion Ratio (SDR)[18]로 원하지 않은 왜곡이 포함된 정도에 대해

평가를 수행하였다. 두 번째는 음성의 지각적인 품질을 평가하기 위해 Perceptual Evaluation of Speech Quality(PESQ)[19]를 이용하였다. PESQ는 ITU-T에서 표준화한 지수로 대부분 1.0에서 4.5 사이의 값을 가진다. 마지막은 음성의 명료도 평가를 위해 Short-Time Objective Intelligibility(STOI)[20] 지수를 이용하였다. STOI 지수는 0에서 1 사이의 값을 가지며, 백분율로 나타내어 성능을 평가하였다.

5.3 실험 결과

실험 결과는 Table 1에서 볼 수 있다. 표에서 'No processing'은 아무 처리도 하지 않은 잡음 오염 음성 데이터를 평가한 결과이다. 'SkipConvNet'은 DCUNET 모델에 SkipConv 모듈만 skip-connection에 적용한 모델로 향상된 음성 데이터에 대한 평가 결과이고 'Skip TFSA Enc'는 DCUNET skip-connection에 TFSA를 적용하도록 제안하는 방법 중 인코더 특징만을 이용하는 방법이다. 또한 'Skip TFSA DE'는 skip-connection에 TFSA를 적용할 때, 인코더와 디코더의 특징을 모두 이용하기 위해 SA에 사용되는 쿼리에 디코더의 특징을 사용한 방법이고 'Skip TFSA Sum'은 인코더와 디코더의 특징을 통합하여 SA를 적용하기 위해 인코더 특징과 디코더 특징을 더한 후에 TFSA를 적용한 방법의 성능 평가 결과이다.

실험 결과, 기존 DCUNET 모델과 비교하여, skip-connection에 TFSA를 적용하였을 때 성능이 최대 SDR 0.17 dB, PESQ 0.13, STOI 1.31 % 향상된 결과를 보인다. 그 중에서 Skip TFSA Enc가 SDR, PESQ 성능지수에서 가장 좋은 성능을 보였고, 이와 비슷하지만 Skip TFSA DE가 STOI 지수에서 가장 좋은 성능을 보였다. Skip TFSA Enc가 가장 좋은 성능을 보이는 것은 SA가 자기 자신의 값만을 이용해 attention을 수행하는 방법이기 때문에 인코더의 특징만을 사용했을 때, 더 효과적으로 적용이 되었다고 볼 수 있다. 또한 Skip TFSA DE에서 skip-connection에 인코더의 특징에 대해 TFSA를 적용할 때, 디코더 특징과의 상관을 이용해 attention을 적용하면서 인코더의 특징이 디코더 특징과 관련되도록 변환되어 음성의 명료도에 더 나은 결과를 보여주었다. 그에 반해 인코더 특징과 디코더 특징을 더하여 TFSA를 수행한 Skip TFSA Sum 방법은 다른 제안하는 방법들과 비교했을 때 성능이 떨어진 결과를 보여준다. 이는 서로 다른 인코더와 디코더 특징을 그대로 더하면서 기존 인코더와 디코더 특징의 형태가 조금 변환되어 성능이 떨어진 것으로 보인다.

Table 1. The result of speech enhancement.

Model	SDR	PESQ	STOI
No processing	0.13	1.587	70.66
UNET	9.97	2.409	85.65
DCUNET	11.32	2.430	86.75
SkipConvNet	11.46	2.544	88.22
Skip TFSA Enc	**11.49**	**2.560**	**88.17**
Skip TFSA DE	**11.46**	**2.558**	**88.26**
Skip TFSA Sum	11.44	2.547	87.93

VI. 결 론

본 논문에서는 DCUNET 모델의 성능 향상을 위해 skip-connection에 SA를 적용하는 방안에 대해 연구를 수행하였다. SA는 높은 복잡성을 줄이고 시간 축, 주파수 축에 따른 상관관계를 모두 고려하기 위해 TFSA를 적용하여 U-Net 구조의 skip-connection에 의해 발생할 수 있는 인코더 특징과 디코더 특징의 불일치 문제를 해소하였다. Skip-connection에 TFSA를 적용하기 위해 skip-connection에서 사용되는 인코더의 특징만을 이용하는 방법과 서로 다른 인코더와 디코더 특징 차이를 줄이기 위해 인코더와 디코더 특징을 TFSA에 이용하는 방안에 대해 연구를 수행하였다. 실험 결과, 기존 DCUNET 모델보다 TFSA를 적용한 방안이 모두 향상된 결과를 보였으며 그 중 skip-connection에 인코더 특징만 사용하는 방법과 인코더 특징과 디코더 특징의 상관관계를 이용하여 SA를 적용한 방안이 가장 좋은 성능을 보였다. 향후에는 Skip TFSA 방법을 이용하여 실제 음성에서 적용해볼 예정이며, 실시간 처리가 가능하도록 구현된 모델에서 향상된 결과를 얻을 수 있는 방안에 대해 연구하고자 한다.

감사의 글

본 논문은 인천대학교 2018년 자체연구비 지원에

의하여 연구되었음.

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,” Proc. EUSIPCO, 1182-1185 (1994).
3. 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).
4. 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).
5. 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,” Proc. ICLR, 1-19 (2018).
6. K. Paliwal, K. Wójcicki, and B. Shannon, “The importance of phase in speech enhancement,” Speech Communication, **53**, 465-494 (2011).
7. Y. Wang and D. L. Wang, “A deep neural network for time-domain signal reconstruction,” Proc. IEEE ICASSP, 4390-4394 (2015).
8. H. Wang, X. Zhang, and D. L. Wang, “Attention-based fusion for bone-conducted and air-conducted speech enhancement in the complex domain,” Proc. IEEE ICASSP, 7757-7761 (2022).
9. 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).
10. V. Kothapally and J. H. Hansen, “Complex-valued time-frequency self-attention for speech dereverberation,” Proc. Interspeech, 2543-2547 (2022).
11. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Proc. NIPS, 6000-6010 (2017).
12. C. Tang, C. Luo, Z. Zhao, W. Xie, and W. Zeng, “Joint time-frequency and time domain learning for speech enhancement,” Proc. 29th IJCAI, 3816-3822 (2021).
13. V. Kothapally, W. Xia, S. Ghorbani, J. H. Hansen, W. Xue, and J. Huang, “Skipconvnet: Skip convolutional neural network for speech dereverberation using optimally smoothed spectral mapping,” Proc. Interspeech, 3935-3939 (2020).
14. M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, “The importance of skip connections in biomedical image segmentation,” Proc. DLMIA, 179-187 (2016).
15. T. Tong, G. Li, X. Liu, and Q. Gao, “Image super-resolution using dense skip connections,” Proc. IEEE ICCV, 4799-4807 (2017).
16. 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).
17. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “Acoustic-phonetic continuous speech corpus CD-ROM NIST speech disc 1-1.1,” DARPA TIMIT, NIST Interagenct/Internal Rep., (NISTIR) 4930, 1993.
18. 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).
19. 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).
20. 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 공학부 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