한국음향학회지 제42권 제6호 pp. 544~551 (2023)
The Journal of the Acoustical Society of Korea Vol.42, No.6 (2023)
https://doi.org/10.7776/ASK.2023.42.6.544

pISSN : 1225-4428
eISSN : 2287-3775

특징 맵 중요도 기반 어텐션을 적용한 복소 스펙트럼 기반 음성 향상에 관한 연구

A study on speech enhancement using complex-valued spectrum employing Feature map Dependent attention gate

정재희,[1] 김우일[1†]

(Jaehee Jung[1] and Wooil Kim[1†])

[1]인천대학교 컴퓨터공학부

(Received August 8, 2023; accepted September 8, 2023)

초 록: 잡음 음성의 지각적 품질과 명료도 향상을 위해 활용되는 음성 향상은 크기 스펙트럼을 이용한 방법에서 크기와 위상을 같이 향상시킬 수 있는 복소 스펙트럼을 이용한 방법으로 연구되어왔다. 본 논문에서는 잡음 음성의 명료도와 품질을 더욱 향상시키기 위해 복소 스펙트럼 기반 음성 향상 시스템에 어텐션 기법을 적용하는 방안에 관해 연구를 수행하였다. 어텐션 기법은 additive attention을 기반으로 수행하며 복소 스펙트럼의 특성을 고려하여 어텐션 가중치를 계산할 수 있도록 하였다. 또한 특징 맵의 중요도를 고려하기 위해 전역 평균 풀링 연산을 같이 사용하였다. 복소 스펙트럼 기반 음성 향상은 Deep Complex U-Net(DCUNET) 모델을 기반으로 수행하였으며, additive attention은 Attention U-Net 모델에서 제안된 방법을 기반으로 연구를 수행하였다. 거실 환경의 잡음 데이터에 대해 음성 향상을 수행한 결과, 제안한 방법이 Source to Distortion Ratio(SDR), Perceptual Evaluation of Speech Quality(PESQ), Short Time Objective Intelligibility(STOI) 평가 지표에서 기준 모델보다 개선된 성능을 보였으며, 낮은 Signal-to-Noise Ratio(SNR) 조건의 다양한 배경 잡음 환경에 대해서도 일관된 성능 향상을 보였다. 이를 통해 제안한 음성 향상 시스템이 효과적으로 잡음 음성의 명료도와 품질을 향상시킬 수 있음을 보여주었다.

핵심용어: 복소 스펙트럼 기반 음성 향상, Additive attention, 특징 맵 기반, 낮은 Signal- to-Noise Ratio (SNR) 환경

ABSTRACT: Speech enhancement used to improve the perceptual quality and intelligibility of noise speech has been studied as a method using a complex-valued spectrum that can improve both magnitude and phase in a method using a magnitude spectrum. In this paper, a study was conducted on how to apply attention mechanism to complex-valued spectrum-based speech enhancement systems to further improve the intelligibility and quality of noise speech. The attention is performed based on additive attention and allows the attention weight to be calculated in consideration of the complex-valued spectrum. In addition, the global average pooling was used to consider the importance of the feature map. Complex-valued spectrum-based speech enhancement was performed based on the Deep Complex U-Net (DCUNET) model, and additive attention was conducted based on the proposed method in the Attention U-Net model. The results of the experiments on noise speech in a living room environment showed that the proposed method is improved performance over the baseline model according to evaluation metrics such as Source to Distortion Ratio (SDR), Perceptual Evaluation of Speech Quality (PESQ), and Short Time Object Intelligence (STOI), and consistently improved performance across various background noise environments and low Signal-to-Noise Ratio (SNR) conditions. Through this, the proposed speech enhancement system demonstrated its effectiveness in improving the intelligibility and quality of noisy speech.

Keywords: Complex-valued spectrum-based speech enhancement, Additive attention, Feature map Dependent, Low Signal- to-Noise Ratio (SNR) environments

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)

Copyright ⓒ2023 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 잡음 음성의 지각적 품질 또는 명료도를 향상시키기 위해 사용되며 다양한 음성 인터페이스 시스템의 성능 향상을 위해 활용된다.

전통적 기법인 스펙트럼 차감법이나 위너 필터와 같은 통계 기반 음성 향상 기법은 잡음이 일정하지 않은 경우나 낮은 Signal- to-Noise Ratio(SNR) 조건의 경우에는 음성 향상 성능에 한계를 가진다.[1,2] 이를 보완하기 위해 심층 신경망 기반 음성 향상의 연구가 수행되어왔다.[3-5] 일반적으로 심층 신경망 기반 음성 향상 기법은 모델을 통해 마스크를 추정하고, 추정된 마스크와 잡음 음성의 스펙트럼을 곱해 향상된 음성의 스펙트럼을 얻을 수 있다.

마스크 기반의 음성 향상 방법은 심층 신경망의 입력으로 사용하는 잡음 음성 특징과 추정된 마스크의 종류에 따라 달라진다.[6] 대표적으로 크기 스펙트럼을 이용한 방법이 있다.[3] 크기 스펙트럼을 이용하는 음성 향상의 경우, 추론 단계에서 향상된 음성의 크기 스펙트럼을 얻은 후 시간 영역의 음성으로 변환할 때 잡음 음성의 위상을 그대로 사용한다. 그러나 크기 스펙트럼뿐만 아니라 위상이 향상된 음성의 품질에 영향을 미친다는 것을 입증하는 연구가 수행되었다.[7,8] 그 결과, 크기와 위상 모두 향상시킬 수 있도록 복소 스펙트럼을 이용한 음성 향상 방법에 관한 많은 연구가 진행되었다.[4,5]

어텐션 메커니즘은 심층 신경망 기반의 연구들에서 많이 활용되었다. 해당 방법은 학습 효과를 높이기 위해 결과에 큰 영향을 미치는 요소에 가중치를 두어, 더 중요한 특징을 집중해서 보기 위해 사용된다. 이를 음성 향상에 적용하는 연구들도 많이 진행되고 있다.[9-12]

본 논문에서는 복소 스펙트럼을 이용한 음성 향상 시스템의 성능 향상을 위해 어텐션을 적용하는 방법에 관한 연구를 수행하였다. 향상된 음성의 스펙트럼은 잡음 스펙트럼과 추정된 마스크에 대해 요소곱을 수행한다. 이런 특성을 고려하기 위해 어텐션은 additive attention[12] 방식을 이용하였다. 어텐션을 적용하여 잡음 스펙트럼 요소 중 깨끗한 음성 관련 특징에 가중치를 두도록 학습한다면 보다 깨끗한 음성에 가깝도록 추정하여 음성의 명료도와 지각적 품질을 향상시킬 것으로 기대할 수 있다. 또한 학습 과정에서 생성되는 모든 특징 맵은 서로 다른 의미와 중요도를 가진다. 이런 특징 맵들 중에서도 비교적 결과에 더 많은 관련이 있는 특징 맵에 가중치를 둔다면 음성 향상 성능에 도움이 될 수 있다. 이를 위해 특징 맵마다 서로 다른 가중치를 적용하는 방안에 대해서도 연구를 수행하였다.

복소 스펙트럼 기반 음성 향상 수행을 위해 기준 모델로 Deep Complex U-Net(DCUNET)[5] 모델의 구조를 이용하였고 Attention U-Net 모델[12]에서 제안된 어텐션 방법을 이용하였다.

먼저, 2장에서는 음성 향상 시스템의 전반적인 수행 과정과 복소 스펙트럼을 이용한 음성 향상 모델에 사용된 복소 연산에 관해 설명한다. 다음으로, 3장에서는 제안하는 additive attention gate와 각 특징 맵을 고려한 attention gate에 대해 설명하고 전체 모델 구조에 관해 설명한다. 마지막으로 4장에서는 사용한 데이터 및 실험에 대해 논의하고, 5장에서 결론을 맺는다.

II. 복소 스펙트럼 기반 음성 향상

2.1 잡음 음성 스펙트럼

음성 향상 모델 훈련 및 추론 기법에서 잡음 음성은 다음과 같이 깨끗한 음성이 시간 축에서 잡음에 부가적으로 오염되는 것으로 모델링된다.

$$y(n) = s(n) + d(n), \tag{1}$$

여기서 y, s, d는 각각 잡음 음성, 깨끗한 음성, 잡음이고 n은 시간 인덱스를 나타낸다. Eq. (1)의 관계는 주파수 영역에서도 유지된다.

$$Y_{t,f} = S_{t,f} + D_{t,f}. \tag{2}$$

Eq. (2)에서 Y, S, D 는 각각 잡음 음성과 깨끗한 음성, 잡음의 스펙트럼을 나타내며, 각각의 스펙트럼은 복소 스펙트럼이다. t는 시간 축에서의 프레임 인덱스를 나타내고, f는 주파수 빈 인덱스를 나타낸다.

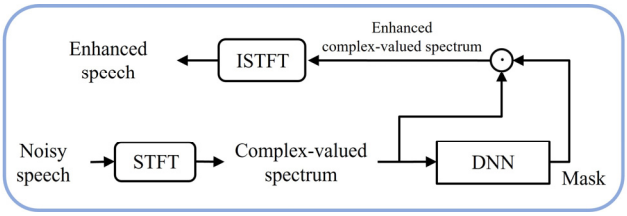

Fig. 1. (Color available online) The process of mask-based speech enhancement system using complex-valued spectrum.

2.2 음성 향상 수행 과정

복소 스펙트럼 기반의 음성 향상 수행 과정은 Fig. 1에서 볼 수 있다. 잡음 음성은 단시간 푸리에 변환(Short Time Fourier Transform, STFT)에 의해 복소 스펙트럼으로 변환되며, 변환된 복소 스펙트럼 Y는 다음과 같이 실수부 Y_r 와 허수부 Y_i로 표현된다.

$$Y = Y_r + jY_i. \tag{3}$$

음성 향상 모델은 잡음 음성의 스펙트럼을 입력으로 사용하여 마스크를 추정한다. 추정된 마스크는 잡음 음성의 스펙트럼과 곱한 후 역 단시간 푸리에 변환을 통해 향상된 음성의 파형을 얻을 수 있다. 본 논문에서는 다음과 같은 연산으로 마스크와 스펙트럼 곱을 수행하였다.

$$\hat{Y} = |Y| \cdot |M| \cdot e^{j(Y_\theta + M_\theta)}, \tag{4}$$

여기서 $\hat{Y}$은 향상된 음성 스펙트럼이며, M은 추정된 마스크이다. $|\cdot|$와 θ는 각각 해당 스펙트럼의 크기와 위상을 나타낸다.

2.3 복소 연산[5]

복소값을 갖는 스펙트럼을 직접 계산할 때는 많은 계산이 필요하므로 복소 스펙트럼 기반의 음성 향상 모델은 다음과 같이 복소 스펙트럼의 실수부와 허수부를 분리한 후 실수 연산으로 변환하여 계산한다. 복소 스펙트럼에 대한 컨볼루션 연산은 다음과 같이 수행된다.

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

$$W * Y = (W_r * Y_r - W_i * Y_i) + j(W_r * Y_i + W_i * Y_r). \tag{6}$$

Eqs. (5)와 (6)에서 $*$은 컨볼루션 연산을 나타내며, W는 복소값 형태의 컨볼루션 필터를 나타낸다. 본 논문에서 사용한 배치 정규화 및 활성화 함수는 컨볼루션 연산과 달리 복소 스펙트럼의 실수 부분과 허수 부분에 각각 적용하였다.

III. Feature map Dependent attention gate

3.1 Additive attention gate

본 논문에서는 음성 손실을 최소화하면서 잡음 음성의 명료도와 지각적인 품질을 향상시키고자 음성 향상 모델에 additive attention을 적용하였다. 어텐션을 이용해 스펙트럼의 요소 중 깨끗한 음성 부분에 가중치를 두고자 하였다. 이를 통해 스펙트럼 요소 중에서 깨끗한 음성에 가까운 부분은 강조되고 그 외 잡음이 존재하는 부분은 상대적으로 억제되어 모델이 깨끗한 음성에 더욱 가깝도록 마스크를 추정하는 데 도움을 줄 수 있다.

Additive attention 방법은 Attention U-Net[12]에서 제안되었던 방법을 기반으로 음성 스펙트럼 특성에 맞게 수정하여 적용하였다. 복소 스펙트럼은 음수와 양수를 포함하고 있어 어텐션 가중치를 계산하는 데에 크기 값을 이용하기 위해 절댓값 연산을 추가로 수행하였고 해당 additive attention gate의 전체적인 구조는 Fig. 2에서 볼 수 있다.

Additive attention gate는 인코더와 디코더의 값을 입력받아 절댓값 연산을 이용해 크기 값으로 변환한다. 이때 위상 정보를 같이 유지해주기 위해 복소값

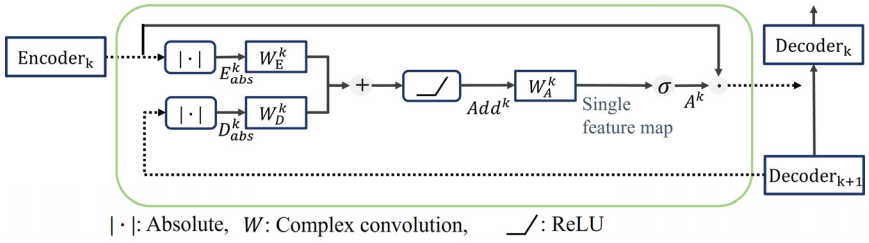

Fig. 2. (Color available online) The proposed additive attention gate.

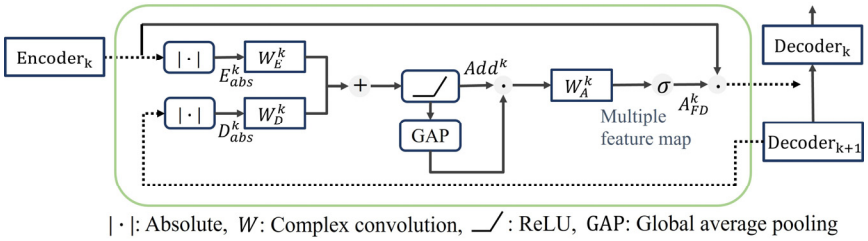

Fig. 3. (Color available online) The proposed Feature map Dependent (FD) attention gate.

행렬 전체가 아닌 실수부, 허수부 각각에 절댓값을 적용하였다. 절댓값 연산을 수행한 후 두 입력값의 정보를 통합하기 위해 컨볼루션을 통한 선형 변환 후에 덧셈 연산을 이용한다. Additive attention gate는 인코더와 디코더의 값을 이용하기 때문에 정보 통합을 통해 압축되면서 잃어버린 정보들을 보완하고 디코더에서 원래 음성의 스펙트럼을 재구성하는 데 도움을 줄 수 있다. 마지막으로 컨볼루션 계층을 통해 특징 맵들을 하나의 특징 맵으로 통합한 후에 시그모이드 활성화 함수를 통해 어텐션 가중치를 추정할 수 있다. 이를 입력으로 사용한 인코더 값에 곱한 뒤 다음 계층 디코더의 입력으로 사용된다.

$$E^k_{abs} = |E^k_r| + j|E^k_i| \in \mathrm{R}^{\mathrm{T_k}\times\mathrm{F_k}\times\mathrm{C_k}}. \quad (7)$$

$$Add^k = ReLU(W^k_E * E^k_{abs} + W^k_D * D^k_{abs}) \in \mathrm{R}^{\mathrm{T_k}\times\mathrm{F_k}\times\mathrm{C_k}}. \quad (8)$$

$$A^k = \sigma(W^k_A * Add^k) \in \mathrm{R}^{\mathrm{T_k}\times\mathrm{F_k}\times 1}. \quad (9)$$

E^k_r과 E^k_i는 각각 k번째 인코더값의 실수부, 허수부로 절댓값 연산을 적용하여 E^k_{abs}를 계산할 수 있다. T_k, F_k, C_k는 각각 k번째 층에 해당하는 특징 맵의 프레임 개수, 주파수 빈 차원, 특징 맵의 개수를 나타낸다. 디코더값 또한 Eq. (7)을 이용해 절댓값을 적용한 디코더 값 D^k_{abs}를 계산할 수 있다. A^k는 k번째 인코더와 디코더값을 이용해 계산한 어텐션 가중치이고, W^k_A, W^k_E, W^k_D는 각각 마지막 컨볼루션 필터, 인코더값에 적용되는 컨볼루션 필터, 디코더값에 적용되는 컨볼루션 필터이다. 여기서, $ReLU$는 Rectified Linear Unit(ReLU) 활성화 함수를 나타내며 σ는 시그모이드 활성화 함수를 나타낸다.

3.2 Feature map Dependent attention gate

3.1에서 제안한 attention gate는 마지막에 어텐션 가중치 추정을 위해 컨볼루션과 시그모이드 활성화 함수를 거치게 된다. 이때 기존 Attention U-Net[12] 모델에서는 채널을 하나로 통합하여 인코더의 특징 맵에 동일한 가중치를 적용하게 된다.

본 논문에서는 각각 다른 의미의 정보와 다른 중요도를 가지는 특징 맵을 고려하기 위해 특징 맵의 중요도 가중치를 계산하여 어텐션을 수행하는 Feature map Dependent(FD) attention gate에 관해 연구를 진행했다. 특징 맵의 중요도를 고려하기 위해 전역 평균 풀링 연산을 적용하였다. 이를 통해 각 특징 맵마다 서로 다른 가중치를 적용하여 학습에 사용할 수 있다. FD attention gate의 전체 구조는 Fig. 3에서 볼 수 있다.

전역 평균 풀링 연산은 인코더와 디코더의 정보

통합 후에 수행되며, 마지막 컨볼루션 층을 거칠 때 특징 맵을 하나로 통합하지 않고 수행한다.

$$A_{FD}^{k} = \sigma(W_{A}^{k} * (Add^{k} \bullet GAP(Add^{k}))) \in \mathrm{R}^{\mathrm{T_k}\times\mathrm{F_k}\times\mathrm{C_k}}. \quad (10)$$

Eq. (10)에서 A_{FD}^{k}는 특징 맵들의 중요도를 고려한 k번째 FD attention gate의 가중치이다. Add^{k}는 Eq. (8)에서 얻은 k번째 attention gate에서 인코더와 디코더의 값이 통합된 결과로 이를 이용해 전역 평균 풀링 연산을 수행한다. 여기서, Global Average Pooling(GAP) 전역 평균 풀링 연산을 나타내며, $GAP(Add^{k})$를 통해 $1\times1\times\mathrm{C_k}$ 형태의 특징 맵 중요도에 따른 가중치를 얻을 수 있다.

3.3 DCUNET with FD attention gate

기존 DCUNET[5]의 모델 구조는 Fig. 4에서 볼 수 있다. DCUNET 모델은 총 8개의 인코더와 디코더로 구성되어 있고 각 인코더와 디코더는 skip-connection을 통해 연결되어 있다. 각 인코더는 컨볼루션, 배치 정규화, 활성화 함수로 구성되어 있으며 디코더는 인코더와 유사하지만 컨볼루션 대신 마스크를 재구성하기 위해 전치 컨볼루션을 수행한다. 마지막 층의 디코더를 제외하고는 모든 층에서 활성화 함수로 Leaky ReLU를 사용하였고, 마지막 층에서는 tanh 함수를 사용하여 마스크의 범위를 -1에서 1로 제한하였다.

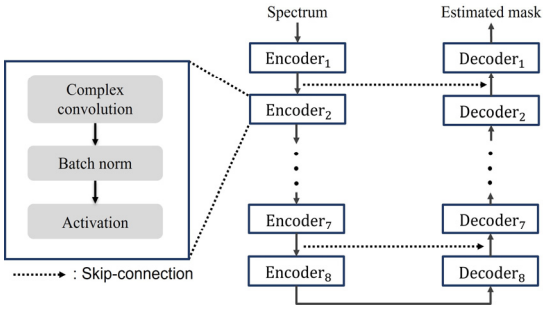

Fig. 4. The baseline system DCUNET.[5]

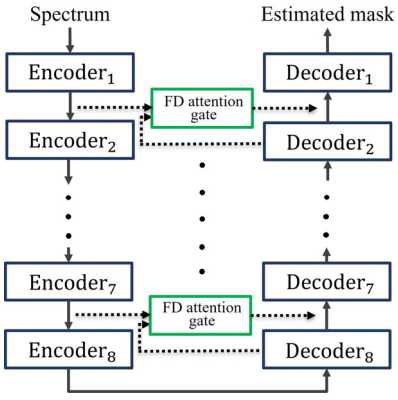

Fig. 5. (Color available online) The proposed DCUNET with FD attention gate.

FD attention gate는 인코더와 디코더가 연결된 skip-connection 사이에 추가하였다. 어텐션을 적용한 모델의 구조는 Fig. 5에서 볼 수 있다. Skip-connection에 attention gate를 적용함으로써 기존에 바로 연결되는 인코더 값보다 중요 특징이 가중된 인코더 값을 연결해줌으로써 더 나은 음성 향상을 수행할 것을 기대할 수 있다.

3.4 손실함수

모델 학습을 위해 SI-SNR(Scale Invariant-Source to Noise Ratio) 손실함수[13]를 사용하였다. SI-SNR 손실은 시간 영역 파형을 이용해 다음과 같이 계산한다.

$$s_{target} := \frac{(<\hat{y}, s> \bullet\ s)}{\|s\|_2^2}. \quad (11)$$

$$e_{noise} := \hat{y} - s_{target}. \quad (12)$$

$$L_{si-snr} := 10\log_{10}\left(\frac{\|s_{target}\|_2^2}{\|e_{noise}\|_2^2}\right). \quad (13)$$

Eq. (11)에서 $<\cdot,\cdot>$은 두 벡터에 대한 내적 곱 연산이고, $\|\cdot\|_2$는 L2 정규화 연산이다. $\hat{y}$은 향상된 음성 파형이고, s는 깨끗한 음성 파형이다.

IV. 실험 및 결과

4.1 데이터베이스

깨끗한 음성 데이터는 TIMIT 데이터베이스[14]를 사용하였고 거실 환경에서 발생하는 TV 잡음을 수집

하기 위해 Youtube에서 TV 프로그램의 오디오 샘플을 수집하여 사용하였다. TV 프로그램 중 드라마, 뉴스, 음악, 스포츠를 이용했으며 잡음은 훈련과 검증, 테스트에서 겹치지 않도록 총 7.5시간의 오디오 파일을 사용하였다. 또한 실제 거실 환경에서 TV의 배경 잡음을 시뮬레이션하기 위해 Room Impulse Response (RIR) 필터를 적용하였다. 잡음 환경은 낮은 SNR을 갖도록 SNR 5 dB, 0 dB, -5 dB 조건으로 깨끗한 음성과 배경 잡음 더해 생성하였다. 생성된 잡음 음성 중 훈련을 위해 55,440 발화가 사용되었고 검증 및 테스트를 위해 1,200 발화와 2,304 발화가 사용되었다.

4.2 실험 설정 및 성능 평가 지표

음성의 샘플링 레이트는 16 KHz이며, 윈도우와 홉 크기는 각각 32 ms와 16 ms로 설정하였다. 단시간 푸리에 변환을 위한 푸리에 변환 개수는 512개로 설정하여 주파수 256차원과 에너지값을 포함하여 사용된 주파수는 총 257차원으로 설정하였다. 훈련을 위해 'Adam' 최적화 알고리즘을 사용하였고 초기 학습률은 0.001로 설정하였다.

성능 평가 비교를 위해 총 3가지 평가 지표를 이용하였다. 첫 번째는 음성 왜곡 정도를 나타내는 Source to Distortion Ratio(SDR)[15]에 대해 평가를 진행하였고 두 번째는 지각적 음성 품질을 평가할 수 있는 Perceptual Evaluation of Speech Quality(PESQ)[16]로 향상된 음성을 평가하였다. 마지막으로는 음성의 명료도 평가를 위해 Short-Time Objective Intelligibility(STOI)[17]를 이용하였다. STOI 값은 0에서 1 사이의 값을 가지며 실험 결과는 백분율로 계산하여 비교를 진행하였다.

또한 'Intel Xeon Gold 5220R CPU @ 2.20 GHz'와 'NVIDIA GeForce RTX 3090 GPU'가 장착된 컴퓨터에서 모델들의 처리 시간을 측정하여 비교하였다.

4.3 실험 결과

실험 결과는 Table 1에서 볼 수 있다. Table 1의 'No processing'은 아무런 처리를 하지 않은 잡음 음성의 평가 결과이다. 기준 시스템으로 크기 스펙트럼을 이용한 음성 향상인 UNET[18]과 복소 스펙트럼을 이용한 DCUNET과의 비교를 진행하였다.

DCUNET+ATT 모델은 DCUNET에 대한 각 특징 맵의 중요도를 고려하지 않은 attention gate를 적용한 모델로 DCUNET에 비해 성능이 향상되었음을 확인할 수 있다. FD attention gate를 적용한 DCUNET + FD-ATT는 DCUNET+ATT에 GAP 연산으로 각 특징 맵에 대한 가중치를 계산하여 어텐션을 적용한 모델로 모든 평가 지표에서 가장 높은 성능을 보여준다.

Tables 2, 3, 4는 각각 SDR, PESQ, STOI에 대한 서로 다른 SNR 조건에서의 평가 결과를 볼 수 있다. 여기서, 결과는 실험에 사용된 배경 잡음인 드라마, 음악, 뉴스, 스포츠에 대한 평균을 계산하였다. 표들을 보면, DCUNET+FD-ATT 모델이 서로 다른 SNR 조건의 잡음 음성에 대해 일관되게 향상된 성능을 보여준다. 이는 본 논문에서 제안하는 FD attention gate가 낮은 SNR 조건의 다양한 배경 잡음 환경에서 음성의 품질 및 명료도를 향상시키는데 상당히 효과적임을 확인할 수 있다.

전체 테스트 데이터에 대해 측정한 처리 시간을 테스트 데이터 개수로 나눈 평균을 계산하여 음성 샘플 당 처리 시간을 비교하였다. 테스트 음성 샘플은 1 s ~ 6 s 사이의 길이이며 평균 길이는 3.04 s이다. 측정 결과, DCUNET은 음성 샘플 당 19.34 ms가 소요되며, DCUNET + ATT와 DCUNET + FD-ATT 모델은 각각 31.37 ms와 36.76 ms가 소요되었다. 이는 FD

Table 1. The evaluation results of the systems. The results are averaged over all test sets.

Model	SDR	PESQ	STOI
No Processing	0.13	1.587	70.66
UNET[18]	9.97	2.409	85.65
DCUNET[5]	11.32	2.430	86.75
DCUNET+ATT	**11.56**	**2.452**	**87.52**
DCUNET+FD-ATT	**11.62**	**2.476**	**87.76**

Table 2. The SDR evaluation results for different SNR conditions. The results are averaged over all types of background noise (drama, music, news, and sports).

Model	SNR5	SNR0	SNR-5	Avg.
No Processing	5.08	0.11	-4.79	0.13
UNET[18]	12.81	9.97	7.13	9.97
DCUNET[5]	13.75	11.40	8.81	11.32
DCUNET+ATT	**14.08**	**11.63**	**8.95**	**11.56**
DCUNET+FD-ATT	**14.19**	**11.72**	**8.96**	**11.62**

Table 3. The PESQ evaluation results for different SNR conditions.

Model	SNR5	SNR0	SNR-5	Avg.
No Processing	1.830	1.563	1.366	1.587
UNET[18]	2.766	2.412	2.048	2.409
DCUNET[5]	2.758	2.444	2.088	2.430
DCUNET+ATT	**2.779**	**2.461**	**2.117**	**2.452**
DCUNET+FD-ATT	**2.801**	**2.492**	**2.136**	**2.476**

Table 4. The STOI evaluation results for different SNR conditions.

Model	SNR5	SNR0	SNR-5	Avg.
No Processing	81.14	70.87	59.98	70.66
UNET[18]	91.37	86.32	79.26	85.65
DCUNET[5]	91.82	87.59	80.86	86.75
DCUNET+ATT	**92.56**	**88.31**	**81.69**	**87.52**
DCUNET+FD-ATT	**92.78**	**88.60**	**81.90**	**87.76**

attention gate를 추가하면서 파라미터가 증가하여 DCUNET 보다 처리 시간이 지연된 결과를 보인다.

V. 결 론

본 논문에서는 잡음 음성의 스펙트럼에서 깨끗한 음성 부분을 보다 효과적으로 강조하여 잡음 음성의 명료도와 지각적 품질을 향상시키기 위해 특징 맵의 중요도를 고려한 attention gate 연구를 수행하였다. FD attention gate는 복소 스펙트럼 특성을 고려하여 어텐션 가중치를 계산하며, 각 특징 맵의 중요도를 고려하기 위해 전역 평균 풀링 방법을 적용하였다. 실험 결과, 제안한 FD attention gate를 적용한 모델이 거실 환경 잡음 데이터에 대해 SDR, PESQ, STOI 평가 지표에서 기준 모델보다 개선된 성능을 보여주었다. 또한 실험한 모든 SNR 조건에서 기준 시스템보다 일관적으로 성능이 향상된 결과를 보여주었다. 이를 통해 제안한 방법이 낮은 SNR 조건의 다양한 배경 잡음 환경에서 잡음 음성의 명료도와 품질을 향상시키는데 상당히 효과적임을 확인하였다. 향후에는 제안한 모델에서 특징 맵의 중요도가 성능에 미친 영향을 면밀히 분석하고 성능을 더욱 향상시킬 수 있는 attention gate 구조에 대해 연구할 예정이다.

감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2021R1F1A1063347).

References

1. J. Lim and A. Oppenheim, “All-pole modeling of degraded speech,” IEEE Trans. Acoust. Speech Signal Process. **26**, 197-210 (1978).
2. S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust. Speech Signal Process. **27**, 113-120 (1979).
3. K. Tan and D. Wang, “A convolutional recurrent neural network for real-time speech enhancement,” Proc. Interspeech, 3229-3233 (2018).
4. 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).
5. 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).
6. D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” IEEE/ACM Trans. Audio, Speech, Language Process. **26**, 1702-1726 (2018).
7. K. Paliwal, K. Wójcicki, and B. Shannon, “The importance of phase in speech enhancement,” Speech Communication, **53**, 465-494 (2011).
8. Y. Wang and D. L. Wang, “A deep neural network for time-domain signal reconstruction,” Proc. ICASSP, 4390-4394 (2015).
9. A. Li, C. Zheng, C. Fan, R. Peng, and X. Li, “A recursive network with dynamic attention for monaural speech enhancement,” Proc. Interspeech, 2422-2426 (2020).
10. Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, “Speech enhancement using self-adaptation and multi-head self-attention,” Proc. ICASSP, 181-185 (2020).
11. Z. Qiquan, S. Qi, N. Zhaoheng, N. Aaron, and L. Haizhou, “Time-Frequency Attention for Monaural Speech Enhancement,” Proc. ICASSP, 7852-7856 (2022).
12. O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, “Attention u-net: Learning where to look for the

pancreas," Proc. MIDL, 1-10 (2018).
13. Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," IEEE/ACM Trans. Audio, Speech, Language Process. **27**, 1256-1266 (2019).
14. 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.
15. E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," IEEE Trans. Audio, Speech, Language Process. **14**, 1462-1469 (2006).
16. 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. ICASSP, 749-752 (2001).
17. C. H. Taal, R. C. Hendriks, and R. Heusdens, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," Proc. ICASSP, 4214-4217 (2010).
18. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," Proc. MICCAI, 234-241 (2015).

저자 약력

▸ 정 재 희 (Jaehee Jung)

2021년 2월 : 인천대학교 컴퓨터공학부 공학사
2021년 3월 ~ 현재 : 인천대학교 컴퓨터 공학과 석사과정

▸ 김 우 일 (Wooil Kim)

1996년 2월, 1998년 8월, 2003년 8월 : 고려대학교 전자공학과 학/석/박사
2004년 8월 ~ 2005년 8월 : Carnegie Mellon University 박사후연구원
2005년 8월 ~ 2012년 8월 : University of Texas at Dallas 연구원, 연구교수
2012년 8월 ~ 현재 : 인천대학교 컴퓨터공학부 조교수, 부교수, 교수