한국음향학회지 제41권 제3호 pp. 342~350 (2022)
*The Journal of the Acoustical Society of Korea* Vol.41, No.3 (2022)
https://doi.org/10.7776/ASK.2022.41.3.342

pISSN : 1225-4428
eISSN : 2287-3775

# 드론 소음 환경에서 심층 신경망 기반 음성 향상 기법 적용에 관한 연구

# A study on deep neural speech enhancement in drone noise environment

김지민,[1] 정재희,[1] 여찬은,[1] 김우일[1†]

(Jimin Kim,[1] Jaehee Jung,[1] Chaneun Yeo,[1] and Wooil Kim[1†])

[1]인천대학교 컴퓨터공학부

(Received March 23, 2022; revised May 10, 2022; accepted May 18, 2022)

**초 록**: 본 논문에서는 재난 환경과 같은 환경에서의 음성 처리를 위해 실제 드론 소음 데이터를 수집하여 오염 음성 데이터베이스를 구축하고 음성 향상 기법인 스펙트럼 차감법과 심층 신경망을 이용한 마스크 기반 음성 향상 기법을 적용하여 성능을 평가한다. 기존의 심층 신경망 기반의 음성 향상 모델인 VoiceFilter(VF)의 성능 향상을 위해 Self-Attention 연산을 적용하고 추정한 잡음 정보를 Attention 모델의 입력으로 이용한다. 기존 VF 모델 기법과 비교하여 Source to Distortion Ratio(SDR), Perceptual Evaluation of Speech Quality(PESQ), Short-Time Objective Intelligibility(STOI)에 대해 각각 3.77 %, 1.66 %, 0.32 % 향상된 결과를 나타낸다. 인터넷에서 수집한 오염 음성 데이터를 75 % 혼합하여 훈련한 경우, 실제 드론 소음만을 사용한 경우에 비해 상대적인 성능 하락률 평균이 SDR, PESQ, STOI에 대해 각각 3.18 %, 2.79 %, 0.96 %를 나타낸다. 이는 실제 데이터를 취득하기 어려운 환경에서 실제 데이터와 유사한 데이터를 수집하여 음성 향상을 위한 모델 훈련에 효과적으로 활용할 수 있음을 확인해준다.
**핵심용어**: 드론 소음, 음성 향상, 스펙트럼 차감법, 마스크 기반 기법, 심층 신경망, Self-Attention

**ABSTRACT:** In this paper, actual drone noise samples are collected for speech processing in disaster environments to build noise-corrupted speech database, and speech enhancement performance is evaluated by applying spectrum subtraction and mask-based speech enhancement techniques. To improve the performance of VoiceFilter (VF), an existing deep neural network-based speech enhancement model, we apply the Self-Attention operation and use the estimated noise information as input to the Attention model. Compared to existing VF model techniques, the experimental results show 3.77%, 1.66% and 0.32% improvements for Source to Distortion Ratio (SDR), Perceptual Evaluation of Speech Quality (PESQ), and Short-Time Objective Intelligence (STOI), respectively. When trained with a 75% mix of speech data with drone sounds collected from the Internet, the relative performance drop rates for SDR, PESQ, and STOI are 3.18%, 2.79% and 0.96%, respectively, compared to using only actual drone noise. This confirms that data similar to real data can be collected and effectively used for model training for speech enhancement in environments where real data is difficult to obtain.
**Keywords:** Drone noise, Speech enhancement, Spectral subtraction, Mask based, Deep neural network, 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

재난 상황과 같이 사람이 직접 구조하기에는 위험하고 한계가 있는 환경에서 드론과 같은 무인기를 이용하는 것이 효과적이다. 재난 환경에서 음향 및 음성 탐지를 위해 드론을 활용하는 요구가 높아지고 있다. 드론으로 사람의 음성을 취득하여 위치를 파악하고자 할 때 드론의 소음은 음성 처리에 방해가 된다.[1] 일반적으로 재난 상황과 같은 특수한 환경에서 인공지능 모델 학습을 위한 충분한 양의 데이터를 구하기는 어렵다.

잡음이 존재하는 환경에서 음성 인식 시스템의 성능을 향상하기 위한 다양한 연구가 진행되어왔다.[2-11] 전통적인 방법으로 위너 필터[2]나 스펙트럼 차감법[4,5]과 같이 음성신호와 잡음 신호의 사이의 통계적 정보를 이용하는 통계 모델 기반의 음성 향상 기법이 오랜 기간 연구되었다. 하지만 시간에 따라 변하는 잡음을 미리 예측하여 제거하는 것이 쉽지 않다는 문제점이 있다. 최근에는 머신러닝 기법인 심층 신경망(Deep Neural Network, DNN)에 관한 연구가 활발히 진행됨에 따라 심층 신경망을 이용한 마스크 기반 음성 향상 기법에 대한 연구가 진행되어 시간에 따라 잡음이 변하는 잡음환경에 대해서도 좋은 성능을 보이는 음성 향상 기법이 연구되고 있다.[7-11]

본 논문에서는 실제 수집한 드론 소음을 이용하여 오염 음성 데이터를 생성하고 통계적 기반의 스펙트럼 차감법과 심층 신경망을 이용한 마스크 기반의 음성 향상 기법을 적용하여 그 성능을 관찰한다. 스펙트럼 차감법에서 잡음 추정 방법으로 통계 최소 기법과 통계 평균 추정 기법을 사용하였다. 심층 신경망을 이용한 마스크 기반 방법에서는 마스크 추정 모델로 VoiceFilter(VF) 모델[7]을 사용하였다. 또한 본 논문은 VF모델의 성능 향상을 위해 VF 모델에 Self-Attention[12-15] 연산을 적용하는 방법을 제안한다. Self-Attention 적용 시에 오염된 음성에서 추정된 잡음을 이용한다. 마스크 추정을 위한 훈련 데이터로 실제 수집 드론 소음을 이용한 음성 데이터만을 이용한 경우, 인터넷 수집 드론 음향을 이용한 음성 데이터만을 이용한 경우, 두 종류의 데이터를 일정 비율로 혼합한 경우에 대해 음성 향상 성능을 비교하였다. 실험을 통해 재난 환경과 같이 실제 드론 소음 데이터를 충분히 확보하지 못하는 환경에서 음성 향상 성능을 위한 데이터를 효과적으로 확보할 수 있는 방안에 대해 타진한다.

본 논문은 다음과 같이 구성된다. 2장에서 실험에 사용된 드론 소음 데이터 수집 및 음성 데이터베이스의 구축 과정에 대해 살펴보고, 3장에서는 실험에 적용한 음성 향상 기법에 대해 설명한다. 4장에서는 실험을 통해 얻은 결과를 비교 및 평가하며, 마지막 5장에서 결론을 맺는다.

# II. 음성 데이터베이스 구축

깨끗한 음성 데이터는 TIMIT[16] 데이터베이스를 사용하였고 드론 소음은 실제 동작하는 드론으로부터 수집한 소음과 유튜브에서 수집한 드론 음향 두 종류로 구성되어 있다. 수집한 드론 소음과 깨끗한 음성 데이터를 3가지 신호대잡음비(Signal-to-Noise Ratio, SNR) 즉, -5 dB, 0 dB, 5 dB의 조건으로 생성하여 드론 소음에 오염된 음성 데이터베이스를 구축하였다.

본 논문의 실험에서는 실제 드론 소음을 이용해 생성한 오염 음성 13,860 샘플, 인터넷 수집 드론 음향을 이용해 생성한 오염 음성 13,860 샘플,[17] 실제 드론 소음 음성과 인터넷 수집 드론 소음 음성을 일정 비율로 혼합한 13,860 샘플을 모델 훈련에 사용하였다. 테스트 데이터는 실제 드론 소음을 이용해 각 신호 대 잡음비 별로 1,680개의 오염 음성 데이터를 생성하여 사용하였다.

## 2.1 실제 드론 소음 데이터 수집

실제 드론 소음은 Fig. 1의 사진과 같이 중형 크기의 드론인 DJI 사의 Air 2S 모델을 소음원으로 사용하여 심장형 지향성을 갖는 스테레오 마이크로폰이 포함된 SONY 사의 PCM-A10 녹음기에 윈드 스크린을 장착한 상태로 측정 환경을 구성하여 수집하였다.

실험에 사용된 드론 소음은 평균 약 75 dB 정도의 크기로 측정되었다. 소음 수집은 드론과 녹음기의 위치에 따라 3가지 경우로 구분하였으며 Fig. 2와 같

이 녹음기를 삼각대에 고정한 상태의 배치 형태로 녹음을 진행하였다. 드론의 위치 및 녹음기와의 거리는 근거리(30 cm) 정지 비행, 중거리(50 cm) 정지 비행, 20 cm에서 60 cm 사이 왕복 수직 이동 비행 3가지로 구성되어 있다.

일정한 드론 소음 데이터를 충분히 얻기 위해 각 비행 상태 당 2 min 씩 5회를 한 세트로 총 5회 녹음을 진행하여 총 150 min가량의 드론 소음 데이터를 수집하였다. 소음 수집은 실제 환경을 고려하여 외부 소음이 존재하는 외부에서 3회 측정하였고 비교 군으로 실내에서 2회 측정하였으나 외부 소음이 실험에 유의미한 영향을 끼칠 정도의 크기가 아니었기에 측정 장소는 구분하지 않고 실험을 진행하였다.

Fig. 1. Photos of the drone and the sound recorder.

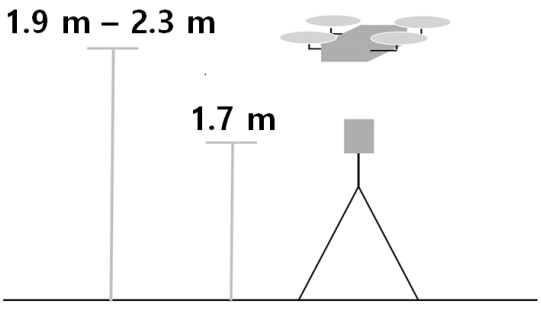



Fig. 2. Layout diagram of the drone and the recorder.

### 2.2 인터넷 수집 드론 음향 데이터

본 연구에서 학습 데이터로 활용하기 위해 다양한 환경에서 다양한 드론으로 녹음한 음향 데이터를 온라인상에서 수집하였다. 실제 녹음한 드론 소음 데이터와는 다른 종류의 드론 음향 데이터를 약 30 min 정도 분량을 수집하여 실험에 사용하였다.

## III. 음성 향상 기법

본 연구에서는 음성 향상 기법으로 기존에 많이 사용되던 통계적 기반의 스펙트럼 차감법과 최근 활발히 연구되고 있는 심층 신경망 기반의 마스크 기반 방법을 사용하였다.

### 3.1 스펙트럼 차감법

스펙트럼 차감법에서 잡음 추정 방법으로 통계 최솟값 추정법과 통계 평균 추정법을 사용하였다. 이는 음성 신호에서 에너지의 최솟값이나 평균값을 잡음으로 추정하여 이를 음성 스펙트럼상에서 차감하는 방법이다. 스펙트럼 차감법의 전체적인 처리 과정은 Fig. 3와 같다.

잡음 음성이 입력으로 들어오면 이산 푸리에 변환을 수행하여 크기와 위상을 얻는다. 스펙트럼의 크기 성분에서 잡음 파워를 추정한 뒤 이를 바탕으로 스펙트럼 가중치를 계산한다. 계산된 가중치를 스펙트럼 크기에 곱해서 향상된 스펙트럼을 얻고 역 이산 푸리에 변환 과정을 통해 향상된 음성을 얻는다.

잡음 파워 추정 과정의 식은 Eqs. (1)과 (2)와 같다. $P_X(\lambda,k)$는 평활화된 입력 신호의 파워값이고 $|X(\lambda,k)^2|$는입력 신호의 파워 스펙트럼이다. $\lambda$는 시간 인덱스, $k$는 주파수 인덱스, $\alpha$는 평활화 상수이다.

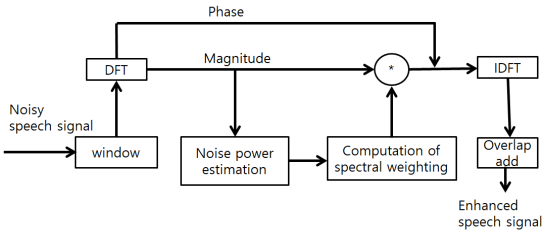



Fig. 3. Process of spectral subtraction method.

Eq. (1)과 같이 평활화된 파워값은 이전 프레임에서 계산한 파워값($P_x(\lambda-1,k)$)과 현재의 파워 스펙트럼의 값($|X(\lambda,k)^2|$)을 합산하여 계산한다. $P_{\min}(\lambda,k)$는 일정 구간 동안 추정된 최소 파워값이고, 보상 계수 $omin$을 곱하여 추정된 잡음 스펙트럼 $P_n(\lambda,k)$ 을 얻는다.

$$P_x(\lambda,k) = \alpha \cdot P_x(\lambda-1,k) + (1-\alpha) \cdot |X(\lambda,k)^2|. \quad (1)$$

$$P_n(\lambda,k) = omin \cdot P_{\min}(\lambda,k). \quad (2)$$

Eq. (2)에서 추정된 잡음을 이용하여 차감을 수행할 때는 Spectral subtraction 논문[4]의 차감 규칙을 따랐다.

## 3.2 마스크 기반 음성 향상 기법

마스크 기반 음성 향상 기법은 음성 스펙트럼의 향상을 표현하는 마스크를 추정하고, 추정된 마스크를 입력 음성 파형이나 스펙트럼에 곱하여 향상된 음성을 얻는 기법이다. 본 논문에서는 마스크를 추정하기 위한 심층 신경망 기반 모델 구조로 VoiceFilter(VF)[7] 모델을 사용하였다. Fig. 4는 VF 모델의 훈련 및 테스트 과정을 도식화 한 것이다. 푸리에 변환 과정을 거친 오염 음성의 스펙트로그램을 VF 모델의 입력으로 넣게 되고 모델의 출력으로 마스크를 얻는다. 훈련 과정에서는 깨끗한 음성의 스펙트로그램과 향상된 스펙트로그램의 차이를 손실함수로 이용한다. 테스트 과정에서는 오염 음성의 스펙트로그램과 추정한 마스크를 곱하여 향상된 스펙트로그램을 얻고 역푸리에 변환 과정을 통해 향상된 음성을 얻는다.

VF 모델은 8개 층의 컨볼루션 신경망(Convolution Neural Network, CNN), 1개 층의 장단기 메모리(Long-Short Term Memory, LSTM), 마지막으로 2개 층의 완전 연결 계층의 모델 구조를 가진다.[7]

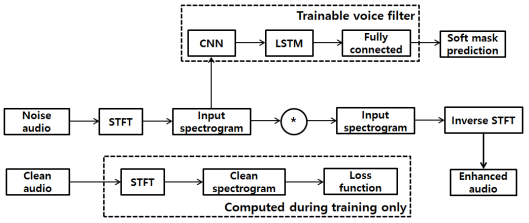



Fig. 4. Training process of mask-based speech enhancement method.

Eq. (3)는 마스크를 추정하는 식이고, Eq. (4)은 사용된 Mean Squared Error(MSE) 손실 함수이다. $|\widehat{M}_{t,f}|$는 마스크, $|\widehat{Y}_{t,f}|$는 오염 음성의 스펙트로그램, $|\widehat{W}_{t,f}|$는 향상된 음성의 스펙트로그램, t는 프레임 인덱스, f와 $F$는 주파수 해상도이다. $G$는 심층 신경망 함수, $|S_{t,f}|$는 깨끗한 음성의 스펙트로그램을 각각 나타낸다.

$$|\widehat{M}_{t,f}| = G(|\widehat{Y}_{t,f}|) \quad (3)$$
$$|\widehat{W}_{t,f}| = |\widehat{Y}_{t,f}| \cdot |\widehat{M}_{t,f}|.$$

$$L_{mse} = \frac{1}{TF}\sum_{t=0}^{T}\sum_{f=0}^{F}(|\widehat{Y}_{t,f}| - |S_{t,f}|)^2. \quad (4)$$

## 3.3 Self-Attention

본 논문에서는 드론 소음 환경에서 음성 향상 성능을 높이기 위해 VF 모델에 Self-Attention[12] 연산을 적용하는 기법을 제안한다. Self-Attention 입력으로는 입력 음성에서 추정된 잡음 정보와 VF 모델의 CNN 출력 값을 이용하고, Attention 결과 값은 LSTM의 입력으로 들어간다. Self-attention 기법에 추정된 잡음과 CNN 출력을 이용함으로써, 잡음에 관련된 특징을 더욱 집중시켜, 음성 향상 진행 시에 Noise-Aware 훈련과 같은 효과를 얻을 수 있을 것으로 기대된다.

Fig. 5는 제안하는 Attention 기법의 과정을 나타낸다. 그림에서 Q, K, V는 각각 Query, Key, Value이다. 입력 음성으로부터 추정된 잡음은 Query로 사용하고 VF 모델의 CNN 층의 출력을 Key와 Value로 사용한다. 전체 Attention 과정은 Eq. (5)로 나타낼 수 있다. Fig. 6 과 같이 Attention 모델의 출력 Z는 VF 모델의 CNN 층의 출력과 연결(concatenation)되어 LSTM 층의 입력으로 들어간다.

$$Z = softmax(\frac{QK^T}{d_k})V. \quad (5)$$

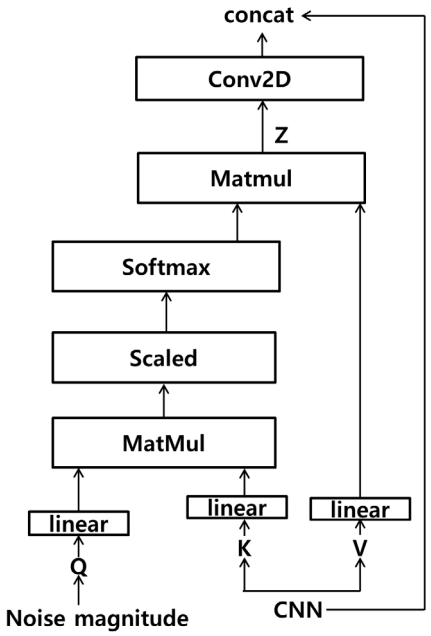



Fig. 5. Architecture of the proposed Self-Attention model.

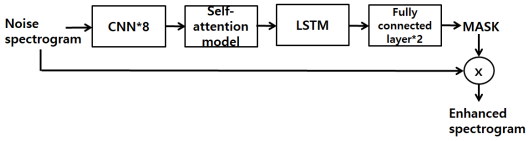



Fig. 6. Block diagram of the proposed speech enhancement system with Self-Attention model.

## IV. 실험 및 결과

### 4.1 데이터베이스 및 실험 환경

깨끗한 음성 데이터는 TIMIT 데이터베이스를 사용하고, 잡음 데이터로는 실제로 수집한 150 min 분량의 드론 소음 데이터와 30 min 분량의 인터넷 수집 드론 소음을 사용하였다. 깨끗한 음성 데이터와 드론 소음 데이터를 SNR 별로 -5 dB, 0 dB, 5 dB의 3가지 조건으로 생성하였다. 마스크 기반 음성 향상 기법의 모델 훈련 과정에서는 실제 수집한 드론 소음으로 생성한 오염 음성 데이터 13,860 샘플과 인터넷에서 수집한 드론 음향으로 생성한 오염 음성 데이터 13,860 샘플을 이용하여 음성 향상 성능을 관찰하고, 실제 드론 소음 오염 음성과 인터넷 드론 소음 음성을 9:1, 4:1, 3:1, 2:1, 1:1, 1:2, 1:3, 1:4, 1:9의 9가지의 비율로 혼합한 훈련 데이터를 이용하여 성능을 비교하였다.

깨끗한 음성과 실제 수집한 드론 소음을 이용하여 3종류의 SNR(-5 dB, 0 dB, 5 dB) 별로 1,680 샘플의 테스트 음성 데이터를 생성하여 음성 향상 기법을 적용하고 그 성능을 관찰하였다. 스펙트럼 차감법에서는 잡음 추정을 위해 통계 최소 추정 기법과 통계 평균 추정 기법을 적용했다. 마스크 기반 음성 향상 기법에서는 실제 드론 소음 오염 음성 데이터와 인터넷 드론 소음 오염 음성 데이터를 각각 훈련 데이터로 사용하여 성능을 비교하였다. 음성 향상 성능 평가는 Source to Distortion Ratio(SDR),[18] Perceptual Evaluation of Speech Quality(PESQ),[19] Short-Time Objective Intelligibility(STOI)[20]를 이용했다.

실험에 사용한 음성 데이터의 샘플링 비율은 8 kHz이며 마스크 기반 기법에서 푸리에 변환을 위해 사용하는 윈도우의 길이는 50 ms, 윈도우의 이동 길이는 20 ms로 설정하였다. FFT의 개수는 512로 설정하였고, 크기 스펙트럼은 에너지 값을 포함하여 총 257 차원으로 설정하여 실험에 사용하였다. 학습률은 0.001로 설정하였고 optimizer로는 adam을 사용하였다.

### 4.2 실제 드론 소음 환경과 인터넷 수집 드론 음향 환경에서의 음성 향상 성능 평가

Table 1은 실험에서 사용한 음성 향상 기법 성능을 각각 SDR, PESQ, STOI로 평가하여 비교한 결과이다. 실험에서의 모든 결과는 3종류의 SNR(-5 dB, 0 dB, 5 dB)에 대한 평균값을 나타낸다. No processing은 아무 음성 향상 기법도 적용시키지 않은 오염 음성에 대한 평가이고, Spsub_avg는 통계 평균값 추정법을 적용한 스펙트럼 차감법, Spsub_min은 통계 최솟값 추정법을 적용한 스펙트럼 차감법을 나타낸다. Mask + Real Drone은 실제 드론 소음에 오염된 음성 데이터를 심층 신경망 모델의 훈련에 사용한 마스크 기반 음성 향상 기법을 나타내고, Mask + Internet Sound는 인터넷에서 수집한 드론 음향 데이터를 이용하여 생성한 오염 음성 데이터를 모델 학습에 사용한 경우를 나타낸다.

실험 결과로부터 마스크 기반 음성 향상 기법이 스펙트럼 차감법과 비교하여 상당히 우수한 음성 향

상 성능을 나타내는 것을 확인할 수 있다. 또한 실제 드론 소음 오염 음성을 이용하여 훈련한 경우(Mask + Real Drone)가 인터넷 드론 음향을 이용하여 훈련한 경우(Mask + Internet Sound)보다 월등히 향상된 결과를 나타내는 것을 확인할 수 있다. 실제 드론 소음을 이용하여 생성한 오염 음성을 테스트 데이터로 사용했기 때문에 테스트 환경과 훈련 환경이 일치하는 Mask + Real Drone 실험이 Mask + Internet Sound에 비해 높은 성능을 내는 것은 당연한 결과로 생각할 수 있다.

인터넷에서 수집한 다양한 환경의 드론 음향을 이용하여 모델을 학습한 경우, 테스트 환경과 일치하는 Mask + Real Drone 실험 결과와 비교할 만한 성능을 나타내는 결과는 인터넷에서 취득한 드론 음향 데이터로 실제 드론 소음 데이터를 마스크 기반 음성 향상을 위한 모델 훈련 용도로 어느 정도 대체할 수 있다는 것을 시사한다.

Table 1. Speech enhancement evaluation results in SDR.

| | SDR | PESQ | STOI |
|---|---|---|---|
| No processing | 0.56 | 2.653 | 0.859 |
| Spsub_avg | 4.05 | 2.839 | 0.855 |
| Spsub_min | 2.78 | 2.700 | 0.850 |
| Mask+Real Drone | 18.81 | 3.673 | 0.934 |
| Mask+Internet Sound | 15.32 | 3.299 | 0.877 |

Table 2. Mask-based speech enhancement evaluation results with various mixed-rates of real drone sound and internet scraped sound for training data.

| Real (%) | Internet (%) | SDR | PESQ | STOI |
|---|---|---|---|---|
| 100 | 0 | 18.81 | 3.673 | 0.934 |
| 90 | 10 | 18.99 | 3.693 | 0.935 |
| 80 | 20 | 18.78 | 3.693 | 0.935 |
| 75 | 25 | 18.79 | 3.681 | 0.934 |
| 67 | 33 | 18.94 | 3.671 | 0.933 |
| 50 | 50 | 18.65 | 3.647 | 0.930 |
| 33 | 67 | 18.37 | 3.643 | 0.930 |
| 25 | 75 | 17.77 | 3.631 | 0.921 |
| 20 | 80 | 18.24 | 3.633 | 0.929 |
| 10 | 90 | 18.20 | 3.625 | 0.928 |
| 0 | 100 | 15.32 | 3.299 | 0.877 |

Table 2는 마스크 기반 음성 향상 기법에서 실제 드론 소음 오염 음성과 인터넷 수집 드론 소음 음성을 다양한 비율로 혼합한 데이터로 훈련한 모델을 이용하여 음성 향상 결과를 얻고 이를 각각의 지표로 평가한 것이다. 예상할 수 있는 것과 같이 인터넷 수집 드론 소음 데이터의 혼합 비율을 높일수록 음성 향상 성능이 점차 하락하여, 인터넷 수집 드론 소음만을 사용할 경우에는 대폭적인 성능 하락을 가져오는 것을 관찰할 수 있다. 하지만 혼합 비율을 1:2 즉, 인터넷 드론 음향 데이터를 67 %로 혼합했을 경우에 실제 드론 소음만을 사용한 경우 대비 성능 하락 비율이 SDR, PESQ, STOI에 대해 각각 2.34 %, 0.82 %, 0.43 %에 불과하고, 혼합 비율을 1:9 즉, 인터넷 드론 음향 데이터를 90 %까지 혼합했을 경우에도 성능 하락 비율이 3.24 %, 1.31 %, 0.64 %에 지나지 않는다. 이와 같은 결과는 재난 상황과 같이 모델 훈련에 충분한 양의 실제 드론 소음을 취득하기 어려운 환경에서는 인터넷 등에서 수집한 드론 음향을 이용하여 음성 향상 모델을 위한 훈련 데이터를 생성하여 사용할 수 있음을 의미한다.

### 4.3 Self-Attention 모델 성능 평가

Table 3은 본 논문에서 제안하는 Self-Attention 기법을 VF 모델에 적용하여 음성 향상 성능을 평가한 결과를 나타낸다. Fig. 5의 구조에서 Query에 해당하는 잡음 정보를 다양한 추정 방법을 사용하여 그 성능을 비교했다. 이 실험에서는 실제 드론 소음 오염 음성 데이터(Real Drone)을 이용하여 모델을 학습했다.

System 1에서는 스펙트럼 차감법의 결과로 얻어지는 깨끗한 음성을 다시 입력 오염 음성에서 차감

Table 3. Speech enhancement evaluation results using Self-Attention model with different types of noise estimation for query of the attention model.

| Query noise information estimation | SDR | PESQ | STOI |
|---|---|---|---|
| System 1: $\widetilde{N}= X-\widetilde{S}$ | 19.34 | 3.705 | 0.937 |
| System 2: Mean($\widetilde{N}$) | 19.09 | 3.711 | 0.937 |
| System 3: None-negative interval mean value | 19.65 | 3.723 | 0.936 |
| System 4: Statistical minimum | 19.52 | 3.734 | 0.937 |

Table 4. Speech enhancement evaluation results using average method for the query noise of the Self-Attention model with various mixed-rates of real drone sound and internet scraped sound for training data.

| Real (%) | Internet (%) | SDR | PESQ | STOI |
|---|---|---|---|---|
| 100 | 0 | 19.65 | 3.723 | 0.936 |
| 75 | 25 | 19.29 | 3.688 | 0.933 |
| 50 | 50 | 18.89 | 3.626 | 0.929 |
| 25 | 75 | 18.81 | 3.626 | 0.928 |

Table 5. Speech enhancement evaluation results using minimum statistics method for the query noise of the Self-Attention model with various mixed-rates of real drone sound and internet scraped sound for training data.

| Real (%) | Internet (%) | SDR | PESQ | STOI |
|---|---|---|---|---|
| 100 | 0 | 19.52 | 3.734 | 0.937 |
| 75 | 25 | 19.15 | 3.675 | 0.933 |
| 50 | 50 | 19.14 | 3.658 | 0.929 |
| 25 | 75 | 18.90 | 3.630 | 0.928 |

하여 얻은 결과를 잡음 정보로 가정하여 Attention 모델의 Query로 사용했다. System 2에서는 System 1에서 얻은 잡음 신호를 평균을 내어 Query로 사용했다. System 3과 System 4에서는 스펙트럼 차감법에서 일반적으로 사용되는 잡음 추정법을 사용했다. System 3에서는 음성의 시작 전 구간과 끝난 후 구간을 비음성 구간으로 가정하여 평균 낸 값을 추정된 잡음으로 사용하는 평균값 추정법을 사용하고, System 4에서는 통계 최솟값 추정법을 사용했다.

실험 결과는 본 논문에서 제안하는 Self-Attention 기법을 VF 모델에 적용한 방법이 Attention을 적용하지 않은 기존의 방법(Table 1의 Mask + Real Drone)과 비교하여 상당한 성능 향상을 가져오는 것을 알 수 있다. 특히 비음성 구간 평균값과 통계 최솟값을 이용하여 추정한 잡음을 Query로 이용한 경우 성능 향상이 컸고, 평균값을 사용한 경우 Table 1의 Mask + Real Drone 결과와 비교하여 SDR이 18.81에서 19.65로 향상되는 것을 알 수 있고, 통계 최솟값을 사용한 경우에는 PESQ는 3.673에서 3.734, STOI는 0.934에서 0.937로 각각 향상되는 것을 알 수 있다. 이러한 결과는 본 논문에서 제안하는 Self Attention 기법에 적절한 추정법을 사용하여 잡음 정보를 Query로 이용하면 잡음에 관련된 특징에 더욱 집중하여 음성 향상에 도움이 되는 것을 입증한다.

Tables 4와 5는 Self Attention 기법을 적용한 모델에서 인터넷 수집 드론 소음 음성 데이터의 활용성을 평가한 결과이다. Table 4는 잡음 추정에 비음성 구간 평균값을 이용한 System 3을 이용한 결과이고, Table 4는 통계 최솟값을 이용한 System 4를 이용한 결과이다.

인터넷 드론 음향 데이터를 50 % 비율로 혼합했을 경우에 System 3은 SDR, PESQ, STOI에 대해 18.89, 3.626, 0.929를 나타내고 System 4는 19.14, 3.658, 0.929를 나타낸다. 이러한 결과는 Table 2에서 나타낸 50 %로 혼합한 VF 모델의 결과인 18.65, 3.647, 0.930와 비교하여 통계 최솟값 잡음 추정과 Attention 기법을 적용한 System 4가 SDR과 PESQ을 향상시키는데 큰 기여를 하고 있음을 나타낸다. 또한 인터넷 드론 음향 데이터를 75 %까지 혼합했을 경우 System 4는 SDR, PESQ, STOI에 대해 18.90, 3.630, 0.928을 나타내고 Table 2의 75 % 혼합 결과인 17.77, 3.631, 0.921의 결과와 비교하여 PESQ를 제외하고 상당히 높은 성능을 보이는 것을 알 수 있다. 또한 75 % 혼합의 경우 실제 드론 소음 음성만을 사용한 경우와 비교하여 성능 하락률이 SDR, PESQ, STOI에 대해 System 3은 4.27 %, 2.61 %, 0.85 %를 나타내고 System 4는 3.18 %, 2.79 %, 0.96 %를 나타내는데, 이는 Table 2의 VF 모델에서 75 % 혼합의 경우 하락률이 5.53 %, 1.14 %, 1.39 % 인 것과 비교하여 PESQ를 제외하고 더 낮은 성능 하락률을 나타내는 것을 확인할 수 있다.

이와 같은 결과는 본 논문에서 제안하는 추정 잡음 정보를 이용한 Attention 기법이 VF 기반의 음성 향상 기술 성능 향상에 효과적임을 입증하고, Attention 기법을 적용한 후에도 인터넷에서 수집한 드론 음향을 이용하여 음성 향상 모델을 위한 훈련 데이터를 생성하여 활용이 가능함을 의미한다.

## V. 결 론

본 논문에서는 재난 환경과 같이 소음이 심한 실제 환경에서의 음성 처리를 위해 실제 드론 소음 데이터

를 수집하여 오염 음성 데이터베이스를 구축하고 음성 향상 기법을 적용하여 성능을 평가했다. 음성 향상 기법으로 스펙트럼 차감법과 심층 신경망을 이용한 마스크 기반 음성 향상 기법을 비교했다. 또한 기존 VF 모델 기반의 음성 향상 기법의 성능을 향상하기 위해 Self-Attention 연산을 적용하고 추정한 잡음 정보를 입력으로 이용하는 기법을 제안했다. 실험 결과는 본 논문에서 제안하는 Self-Attention 기법을 적용한 VF 모델의 음성 향상 평가 지표 SDR, PESQ, STOI는 19.52, 3.734, 0.937로 기존의 VF 모델과 비교하여 모든 지표가 상당히 향상된 결과를 보여주었다. 이와 같은 결과는 본 논문에서 제안하는 Attention 기법이 입력 음성 신호의 잡음 관련 특징에 보다 집중함으로써 음성 향상에 도움을 주는 것을 입증한다. 인터넷에서 수집한 드론 음향을 이용한 오염 음성 데이터를 75 % 혼합하여 훈련 데이터에 사용한 경우 SDR, PESQ, STOI는 18.90, 3.630, 0.928를 나타내고 실제 드론 소음만을 사용한 경우에 비해 상대적인 성능 하락률이 각각 3.18 %, 2.79 %, 0.96 %을 나타냈다. 이러한 결과는 재난 환경과 같이 실제 드론 소음을 취득하기 어려운 환경에서 실제 데이터와 유사한 데이터를 인터넷 등에서 수집하여 음성 향상을 위한 모델 훈련에 효과적으로 활용할 수 있는 가능성을 확인해준다. 향후 연구에서는 실제 드론에 탑재된 마이크를 이용하여 음성을 수집하여 실험에 사용함으로써 실제 상황에 가까운 재난 소음 환경에서의 음성 향상 실험을 진행할 계획이다. 또한 인터넷 수집 소음 데이터의 활용을 극대화하기 위해 데이터 보강(Augmentation), 전이(Transfer) 학습 등을 적용하여 성능을 비교하고자 한다.

## 감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2019R1F1A106299513).

## References

1. M.Narinen, *Active noise cancellation of drone propeller noise through waveform approximation and Pitch-shifting*, (Ph.D. thesis, Georgia State University, 2020).
2. J. Lim and A. Oppenheim, “All-pole modeling of degraded speech,” IEEE Trans. on Acoustics, Speech, and Signal Process. **26**, 197-210 (1978).
3. Y. Ephraim and D. Malah, “Speech enhancement using minimum mean square error short time spectral amplitude estimator,” IEEE Trans. on Acoustics, Speech and Signal Process. **32**, 1109-1121 (1984).
4. S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. on Acoustics, Speech and Signal Proccess. **27**, 113-120 (1979).
5. R. Martin, “Spectral subtraction based on minimum statistics,” Proc. EUSIPCO, 1182-1185 (1994).
6. P. J. Moreno, B. Raj, and R. M. Stern, “Data-driven environmental compensation for speech recognition: a unified approach,” Speech Communication, **24**, 267-285 (1998).
7. Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, “Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking,” arXiv preprint arXiv:1810.04826 (2018).
8. C. Deng, H. Song, Y. Zhang, Y. Sha, and X. Li, “DNN-based mask estimation integrating spectral and spatial features for robust beamforming,” Proc. IEEE ICASSP, 4647-4651 (2020).
9. N. Saleem, M. I. Khattak, M. Al-Hasan, and A. B. Qazi, “On learning spectral masking for single channel speech enhancement using feedforward and recurrent neural networks,” in IEEE Access, **8**, 160581-160595 (2020).
10. M. Hasannezhad, Z. Ouyang, W. -P. Zhu, and B. Champagne, “Speech enhancement with phase sensitive mask estimation using a novel hybrid neural network,” IEEE Open Journal of Signal Processing, **2**, 136-150 (2021).
11. M. Hasannezhad, Z. Ouyang, W. -P. Zhu, and B. Champagne, “An integrated CNN-GRU framework for complex ratio mask estimation in speech enhancement,” Proc. APSIPA ASC, 764-768 (2020).
12. Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, “Speech enhancement using self-adaptation and multi-head self-attention,” Proc. IEEE ICASSP, 181-185 (2020).
13. X. Hao, C. Shan, Y. Xu, S. Sun, and L. Xie, “An attention-based neural network approach for single channel speech enhancement,” Proc. IEEE ICASSP, 6895-6899 (2019).
14. S. K. Roy, A. Nicolson, and K. K. Paliwal, “Deep LPC-MHANet: Multi-head self-attention for augmented

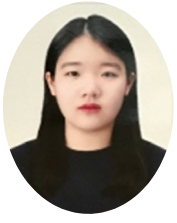
kalman filter-based speech enhancement," IEEE Access, **9**, 70516-70530 (2021).
15. A. Pandey and D. Wang, "Dense CNN with self-attention for time-domain speech enhancement," IEEE/ACM Trans Audio Speech Lang Process. **29**, 1270-1279 (2021).
16. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA STI/Recon Tech. Rep. 1993.
17. *Syma X5C-1*, https://youtube.com/watch?v=aR3NgjOwzAo&featur e=share, (Last viewed August 19, 2021).
18. E. Vincent, R. Gribonval, and C. Fevotte, "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, 01CH37221 (2001).
20. C. H. Taal, R. C. Hendriks, R, Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time- frequency weighted noisy speech," Proc. IEEE ICASSP, 4214-4217 (2010).

## 저자 약력

▸ 김 지 민 (Jimin Kim)

2021년 3월 ~ 현재 : 인천대학교 컴퓨터 공학부 학사과정

▸ 정 재 희 (Jaehee Jung)



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

▸ 여 찬 은 (Chaneun Yeo)

2019년 3월 ~ 현재 : 인천대학교 컴퓨터 공학부 학사과정

▸ 김 우 일 (Wooil Kim)

1996년 2월, 1998년 8월, 2003년 8월 : 고려대학교 전자공학과 학/석/박사
2012년 8월 ~ 현재 : 인천대학교 컴퓨터공학부 조교수, 부교수, 교수