JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 24, No. 4: 459~466, Apr. 2020

멀티로터 UAV 환경에서의 CNN 기반 복소 스펙트로그램 향상 기법

김영진[1] · 김은경[2*]

CNN based Complex Spectrogram Enhancement in Multi-Rotor UAV Environments

Young-Jin Kim[1] · Eun-Gyung Kim[2*]

[1]Ph.D. student, Department of Computer Science & Engineering, Korea University of Technology and Education, Cheonan, 31253 Korea
[2*]Professor, School of Computer Science & Engineering, Korea University of Technology and Education, Cheonan, 31253 Korea

요 약

멀티로터 UAV(Unmanned Aerial Vehicle)를 이용해서 수집한 음향 데이터는 모터나 프로펠러에서 발생하는 자체 소음이나 비행 중 발생하는 바람 소리 등으로 인해 음향 품질이 크게 손상되는 문제가 발생한다. 멀티로터 UAV 환경에서는 목표 음향의 크기뿐만 아니라 위상도 크게 손상되기 때문에 크기와 위상을 모두 고려해서 음향을 향상시킬 필요가 있다. 하지만 위상은 크기와 달리 구조적인 특징이 잘 나타나지 않으므로 향상시키는 것이 쉽지 않다. 따라서 본 연구에서는 크기와 위상을 모두 표현할 수 있는 복소 스펙트로그램을 기초로 잡음을 제거해서 목표 음향의 품질을 향상시키는 CNN 기반 복소 스펙트로그램 향상 방법을 제안한다.

ABSTRACT

The sound collected through the multi-rotor unmanned aerial vehicle (UAV) includes the ego noise generated by the motor or propeller, or the wind noise generated during the flight, and thus the quality is greatly impaired. In a multi-rotor UAV environment, both the magnitude and phase of the target sound are greatly corrupted, so it is necessary to enhance the sound in consideration of both the magnitude and phase. However, it is difficult to improve the phase because it does not show the structural characteristics. in this study, we propose a CNN-based complex spectrogram enhancement method that removes noise based on complex spectrogram that can represent both magnitude and phase. Experimental results reveal that the proposed method improves enhancement performance by considering both the magnitude and phase of the complex spectrogram.

키워드 : 음향 향상, UAV, 딥 러닝, CNN, 음향 신호 처리

Keywords : Sound Enhancement, UAV, Deep Learning, Convolutional Neural Network, Acoustic Signal Processing

Received 28 January 2020, **Revised** 14 February 2020, **Accepted** 1 March 2020

*** Corresponding Author** Eun-Gyung Kim(E-mail: egkim@koreatech.ac.kr Tel:+82-41-560-1350)
Professor, School of Computer Science & Engineering, Korea University of Technology and Education, Cheonan, 31253 Korea

Open Access http://doi.org/10.6109/jkiice.2020.24.4.459 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

최근 음향 센서를 부착한 멀티로터 UAV(Unmanned Aerial Vehicle)를 활용해서 공중이나 지상에서 발생한 음향을 수집하고, 처리 및 분석하는 방법에 대한 연구가 활발하다. 특히, 멀티로터 UAV 음향 시스템을 이용하면 항공 영상에서 청각적 콘텐츠를 고려할 수 있으며, 구조 음성의 방향 추정을 통한 실종자 수색 및 인명 구조, 감시 등의 용도로 사용할 수 있다는 장점이 있다[1, 2, 3]. 하지만 멀티로터 UAV의 모터와 프로펠러에서 발생하는 자체 잡음과 비행 중 발생하는 바람 소리 등은 목표 음향의 품질을 급격히 떨어뜨려 음향의 처리 및 분석을 어렵게 하는 문제가 있다. 자체 잡음은 모터의 회전에 따라 발생하는 잡음과 프로펠러에 공기가 부딪치면서 발생하는 마찰음 등으로 구성되며, 모두 시간에 따라 불안정적(non-stationary)인 특성을 갖는다. 또한, 자체 잡음은 주파수 영역 전반에 걸쳐 분포하기 때문에 목표 음향의 전반적인 주파수 영역을 손상시키는 문제를 야기한다. 바람 소음은 멀티로터 UAV가 비행함에 따라 직/간접적으로 마이크에 전달되는 소음으로, 비행 속도나 바람의 세기에 따라 크게 변화하는 불안정적인 특성을 갖는다. 또한, 바람 소음은 낮은 주파수 영역에 주로 분포하기 때문에 음향 신호의 주요 정보가 담겨 있는 원천 주파수(fundamental frequency) 신호의 품질을 크게 떨어뜨린다. 따라서 멀티로터 UAV에서 수집한 음향은 불안정적이고 강한 잡음들로 인하여 직접적인 처리 및 분석이 쉽지 않으므로, 음향 향상(Sound Enhancement) 기술을 적용해서 자체 잡음과 바람 소음을 제거할 필요가 있다.

기존의 신호처리 및 확률 모델 기반의 음향 향상 방법으로는 spectral subtraction[4]과 wiener filtering[5], minimum mean-squared error(MMSE)[6]가 있다. 그러나 이런 방법들은 잡음 및 소음이 짧은 시간 구간에서 크게 변화하지 않는 안정적(stationary)이라는 가정 아래 통계적 특징이나 확률 모델을 만들기 때문에, 멀티로터 UAV의 자체 잡음과 바람 소음처럼 매우 불안정적인 잡음에 의해 변질된 음향은 잘 향상시키지 못하는 단점이 있다. 반면, 최근 부상하고 있는 데이터 지향(data-driven)의 딥러닝[7, 8, 9] 기반 음향 향상 기술은 신호처리 및 확률 모델 기반 음향 향상 기술의 성능을 크게 능가하고 있다. 딥러닝 기반 음향 향상 기법은 학습 대상에 따라 크게 T-F(Time-Frequency) Mask 기반 음향 향상 기법[10, 11, 12]과 Direct Mapping 기반의 음향 향상 기법[13, 14, 15]으로 나눌 수 있다. T-F Mask 기반 음향 향상 연구인 [10, 11, 12]에서는 잡음이 포함된 음향의 시간-주파수 표현인 Mixture 스펙트로그램(spectrogram)으로부터 T-F Mask를 추정하도록 딥러닝 모델을 학습시켰으며, 추론 단계에서는 추정된 T-F Mask와 Mixture 스펙트로그램을 곱해서(element-wise multiplication) 음향을 향상시켰다. Direct Mapping 기반의 음향 향상 연구인 [13, 14, 15]에서는 Mixture 스펙트로그램을 직접 딥러닝 모델에 전달함으로써 모델을 학습시키고, 목표 음향 스펙트로그램을 추정해서 음향을 향상시켰다. 그러나 연구 [10, 13, 14]과 같은 대부분의 음향 향상 방법들은 크기(magnitude) 스펙트로그램이나 그 변형인 LPS(Log Power Spectra)의 향상에만 초점을 맞추고 있다. 시간-주파수 표현인 스펙트로그램을 시간 도메인의 음향 신호로 재구성하기 위해서는 개별 주파수 신호들에 대한 크기와 위상(Phase)이 모두 필요하다. 따라서 연구 [10, 13, 14]과 같이 크기 스펙트로그램만 향상시킬 경우, 잡음이 포함된 Mixture 음향의 위상 스펙트로그램을 추가로 이용해서 시간 도메인 음향 신호로 재구성해야 한다. 반면, 잡음이 목표 음향보다 지배적인 환경(low SNR)에서는 Mixture 음향의 위상 스펙트로그램과 목표 음향의 위상 스펙트로그램 간에 큰 차이가 발생하므로, 크기 스펙트로그램을 잘 향상시키더라도 시간 도메인 신호의 재구성 과정에서 큰 오차가 발생한다. 연구 [11]에서는 시간 도메인의 신호를 재구성하는 과정에서 발생하는 오차를 최소화하기 위해 수식 (1)과 같은 PSM(Phase Sensitive Mask)을 제안하였다.

$$M_{PSM}(t,f) = \frac{|s(t,f)|}{|x(t,f)|} cos\theta \tag{1}$$

수식 (1)의 $|s(t,f)|$는 목표 음향의 크기 스펙트로그램을 의미하고, $|x(t,f)|$는 Mixture 크기 스펙트로그램을 의미한다. 또한, $\cos\theta$는 위상 간 오차를 보상하기 위한 코사인 항이다. 즉, PSM은 크기 스펙트로그램에 대한 Mask에 신호 재구성 과정에서 발생하는 위상 간 오차를 보상할 수 있는 코사인 항을 추가함으로써 효과적으로 음향을 향상시켰다. 그러나 PSM 기반 음향 향상 또한 잡음이 매우 지배적인 음향에서의 위상 오차를 완전하게 제거할 수 없다는 문제가 있다. 연구 [12]에서는 시간

-주파수 도메인의 음향을 직교 좌표계에서의 실수부와 허수부로 표현하였으며, 수식 (2)와 같이 실수부와 허수부에 모두 적용할 수 있는 cIRM(complex Ideal Ratio Mask)을 제안하였다.

$$M_{cIRM} = \frac{s(t,f)}{x(t,f)} = \frac{x_r(t,f)s_r(t,f)+x_i(t,f)s_i(t,f)}{x_r^2(t,f)+x_i^2(t,f)} + j\frac{x_r(t,f)s_i(t,f)-x_i(t,f)s_r(t,f)}{x_r^2(t,f)+x_i^2(t,f)} \quad (2)$$

수식 (2)의 x_r, x_i은 각각 Mixture 복소 스펙트로그램의 실수부와 허수부를 의미하고, s_r, s_i는 각각 목표 음향 복소 스펙트로그램의 실수부와 허수부를 의미한다. 시간-주파수 도메인의 직교 좌표계 표현인 실수부와 허수부를 모두 향상시킬 경우, 극 좌표계에서의 크기와 위상을 모두 향상시키는 것과 동일한 결과를 나타낼 수 있으므로, cIRM을 통해 향상된 복소 스펙트로그램은 시간 도메인으로의 완전한 재구성이 가능하다. 반면, cIRM은 값의 범위가 (-∞,∞)로 매우 넓기 때문에 목표 음향에 대한 cIRM 추정이 매우 어렵다는 문제가 있다.

Direct Mapping 기반의 연구 [15]에서는 cIRM과 유사하게 복소 스펙트로그램의 실수부와 허수부를 직접 향상시키는 CNN(Convolutional Neural Network)을 제안하였다. 연구 [15]의 음향 향상 방법은 복소 스펙트로그램의 실수부와 허수부를 직접 향상시키기 때문에, 추정하고자 하는 값의 범위가 크지 않으므로 효과적인 음향 향상이 가능하다. 반면, 복소 스펙트로그램의 실수부와 허수부는 서로 일정한 관계를 나타냄에 반해, 연구 [15]의 딥러닝 구조는 이러한 관계를 고려하지 않는다는 문제가 있다.

따라서 본 연구에서는 멀티로터 UAV와 같이 강한 잡음 환경에서 크기와 위상을 모두 고려할 수 있는 복소 스펙트로그램을 기초로, 실수부와 허수부 사이의 관계를 고려하면서 음향을 향상시킬 수 있는 CNN 모델인 CSEN(Complex Spectrogram Enhancement Network)을 제안한다.

Ⅱ. 강한 잡음환경에서의 위상

음향 신호의 시간-주파수 표현은 극 좌표계에서 크기와 위상, 직교 좌표계에서 실수부와 허수부로 나타낼 수 있다. 또한, 극 좌표계에서의 크기와 위상은 각각 수식 (3)과 (4)처럼 실수부와 허수부에 대한 식으로 표현할 수 있다.

$$|s_{t,f}| = \sqrt[2]{\mathcal{R}(s_{t,f})^2+\mathfrak{I}(s_{t,f})^2} \quad (3)$$

$$\theta_{s_{t,f}} = \tan^{-1}\frac{\mathfrak{I}(s_{t,f})}{\mathcal{R}(s_{t,f})} \quad (4)$$

수식 (3)의 $|s_{t,f}|$는 크기 스펙트로그램을 의미하고, 수식 (4)의 $\theta_{s_{t,f}}$는 위상을 의미하며, $\mathcal{R}$,$\mathfrak{I}$는 각각 실수부와 허수부를 의미한다.

음향 신호의 시간-주파수 표현으로의 변환은 일반적으로 STFT(Short Time Fourier Transform)을 사용하는데, STFT는 선형적 특성을 가지므로 시간 도메인 음향 신호는 시간-주파수 도메인에서 수식 (5)와 같이 표현할 수 있다.

$$\begin{aligned} x_{t,f} &= s_{t,f}+n_{t,f} \\ &= |s_{t,f}|e^{\theta_{s_{t,f}}}+|n_{t,f}|e^{\theta_{n_{t,f}}} \\ &= (\mathcal{R}(s_{t,f})+\mathcal{R}(n_{t,f}))+i(\mathfrak{I}(s_{t,f})+\mathfrak{I}(n_{t,f})) \end{aligned} \quad (5)$$

이때, Mixture 신호의 위상은 수식 (6)과 같이 나타낼 수 있다.

$$\theta_{x_{t,f}} = \tan^{-1}\frac{\mathfrak{I}(s_{t,f})+\mathfrak{I}(n_{t,f})}{\mathcal{R}(s_{t,f})+\mathcal{R}(n_{t,f})} \quad (6)$$

수식 (6)과 같이 Mixture 신호의 위상에 나타난 결과로 미루어 볼 때, 잡음이 목표 음향보다 강하지 않은 환경($\mathcal{R}(s_{t,f}) \gg \mathcal{R}(n_{t,f}),\quad \mathfrak{I}(s_{t,f}) \gg \mathfrak{I}(n_{t,f})$)에서는 Mixture 위상 스펙트로그램을 목표 위상 스펙트로그램의 위상으로 근사할 수 있으므로 시간 도메인 음향 신호의 재구성에서 큰 오차가 발생하지 않는다. 반면, 잡음이 목표 음향보다 지배적인 환경에서는 Mixture 위상 스펙트로그램이 목표 위상 스펙트로그램을 나타내지 못하기 때문에, 시간 도메인 음향 신호로의 재구성 과정에서 큰 오차가 발생한다.

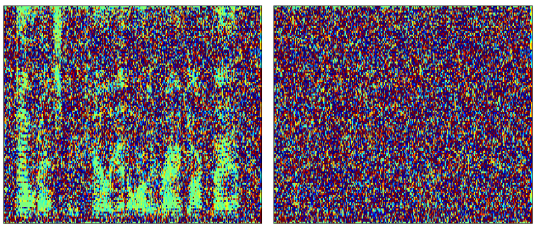

Fig. 1 Difference between Mixture Phase and Target Phase at different SNR

그림 1은 약한 잡음 환경(좌측, SNR>15dB)과 강한 잡음 환경(우측, SNR<-15dB)에서의 Mixture 신호의 위상 스펙트로그램과 목표 신호의 위상 스펙트로그램의 차이를 나타낸 것이다. 그림 1의 연녹색으로 표현된 영역은 Mixture 신호의 위상 스펙트로그램과 목표 신호의 위상 스펙트로그램의 차이가 0에 가까운 영역을 의미하며, 그림 1의 좌측과 같이 잡음이 목표 음향보다 강하지 않은 상황에서는 Mixture 신호의 위상이 목표 신호의 위상을 나타내는 것을 알 수 있다. 반면, 그림 1의 우측과 같이 잡음이 지배적인 상황에서는 Mixture 신호의 위상이 목표 신호의 위상을 나타내지 못한다. 따라서 멀티로터 UAV와 같이 강한 잡음 환경에서는 크기 스펙트로그램뿐만 아니라 위상 스펙트로그램 또한 향상시킬 필요가 있다.

Ⅲ. 복소 스펙트로그램의 실수부와 허수부 간의 관계

직교 좌표계 표현인 복소 스펙트로그램의 실수부와 허수부는 오일러 공식(Euler's formula)에 따라 수식 (7)과 같이 극 좌표계 표현인 크기와 위상의 조합으로 표현된다. 따라서 복소 스펙트로그램의 실수부와 허수부를 모두 향상시킬 경우, 극 좌표계에서의 크기와 위상 스펙트로그램을 모두 향상시키는 것과 동일한 결과를 얻을 수 있다.

$$\mathcal{R}(s_{t,f}) = |s_{t,f}|\cos(\theta_{s_{t,f}}) \qquad (7)$$
$$\mathfrak{I}(s_{t,f}) = |s_{t,f}|\sin(\theta_{s_{t,f}})$$

이때, 삼각함수인 sin과 cos은 삼각함수의 항등성에 따라 수식 (8)와 같은 관계로 나타나기 때문에, 복소 스펙트로그램의 실수부와 허수부는 그림 2와 같이 서로 매우 유사한 구조적 특징을 갖는다.

$$|s_{t,f}|\cos(\theta_{s_{t,f}}) = |s_{t,f}|\sin(\frac{\pi}{2} - \theta_{s_{t,f}}) \qquad (8)$$
$$|s_{t,f}|\sin(\theta_{s_{t,f}}) = |s_{t,f}|\cos(\frac{\pi}{2} - \theta_{s_{t,f}})$$

그림 2는 로그-스케일로 변환한 복소 스펙트로그램의 실수부와 허수부를 시각화한 것이다.

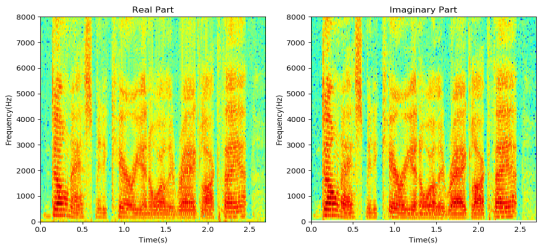

Fig. 2 Log-Scaled Real & Imaginary Component of Complex Spectrogram

복소 스펙트로그램의 실수부와 허수부는 위상과 달리 구조적 특징이 잘 나타나기 때문에 향상시키는 것이 비교적 쉬울 뿐만 아니라, 실수부와 허수부를 모두 향상시키면 크기와 위상을 모두 향상시키는 것과 같은 결과를 얻을 수 있다는 장점이 있다. 또한, 실수부와 허수부는 그림 2와 같이 서로 유사한 구조를 나타내기 때문에 단일 필터를 공동으로 적용해서 향상시킬 수 있다. 반면, 실수부와 허수부는 서로 위상이 $\pi/2$만큼 이동된 형태로 차이가 발생하기 때문에 실수부와 허수부 간의 관계를 고려할 필요가 있다. 따라서 본 연구에서는 실수부 허수부 모두에 대해 공유된(공통된) 단일 필터를 공동으로 적용해서 실수부와 허수부를 향상시키면서, 채널 간 weighted linear transform을 통해 실수부와 허수부 사이의 관계를 반영해서 음향을 향상시키는 방법을 제안한다.

Ⅳ. 복소 스펙트로그램 향상을 위한 CNN 기반의 CSEN 설계 및 구현

본 연구에서는 마이크로부터 수신된 음향 신호를 STFT를 이용해서 복소 스펙트로그램으로 변환한 후, 복소 스펙트로그램의 실수부와 허수부들을 별도의 채널을 기준으로 연결해서 $T \times F \times C$ 크기의 입력 특징

을 구성하였다. T는 시간 축, F는 주파수 축, 그리고 C는 채널 축을 의미한다. 이후, 구성된 입력 특징은 본 연구에서 제안한 CSEN을 통과해서, $F\times C$ 크기의 목표 음향에 대한 출력 특징을 추정한다. CSEN은 Direct Mapping 기반의 음향 향상 모델로서, 개념적 구성은 그림 3과 같다.

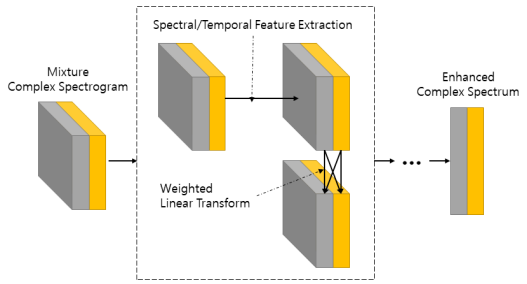

Fig. 3 Conceptual Architecture of Proposed Enhancement Network

CSEN의 주요 특징은 시간-주파수 도메인에 대한 필터와 실수부와 허수부로 구성된 채널에 대한 필터를 구분해서 적용하는 것이다. 즉, $T\times F\times C$의 입력 특징에 대해, 하나의 공유된 필터를 적용해서 시공간적 특징을 추출한 후, 채널에 대한 weighted linear transform을 적용해서 실수부 허수부 간 관계를 반영한다.

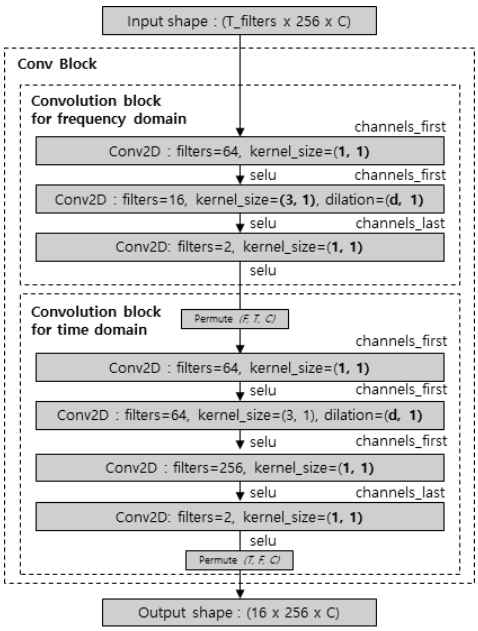

Fig. 4 Configuration of Conv. Block

본 연구에서는 시공간적 특징 추출을 위한 Backbone Network로 이전 연구 [16]에서 활용했던 DCN(Densely Connected Networks)[14]을 수정해서 사용하였으며, 주요 특징 추출 계층인 Conv. Block을 그림 4와 같이 구성하였다.

본 연구에서 제안한 CSEN은 각 실수부와 허수부에 대해 공유된 필터를 적용하기 때문에, 필요한 파라미터 수가 많지 않다는 장점이 있다.

Ⅴ. 실험 및 결과 분석

5.1. 데이터 셋 구성

목표 음향에 해당하는 음성 데이터는 TIMIT 데이터셋[17]을 사용했다. TIMIT는 다양한 성별과 방언을 갖는 630명의 미국인 화자가 발성한 총 6,300개의 음성 파일로 구성되어 있다. 본 연구에서는 TIMIT 음성 파일들을 8:1:1의 비율로 학습, 검증, 테스트 셋을 구성하였다.

멀티로터 UAV의 자체 잡음은 Sunnysky X4108S KV380 모터와 DUALSKY 13x5.5 카본 프로펠러가 부착된 멀티로터 UAV에 멀티로터 UAV 자체 기준 0.1m 아래에 마이크를 부착하여 수집하였다. 잡음 음향 수집을 위해 멀티로터 UAV에 부착한 마이크는 Sony PCM-A10 레코드를 사용하였다. 학습에 사용된 멀티로터 UAV 잡음 음향은 자유 비행하는 멀티로터 UAV에서 약 30분 동안 수집하였으며, 검증 및 테스트에 사용된 멀티로터 UAV 잡음은 hovering에서 3분, shifting에서 3분, rotating에서 2분 동안 수집하였고, 각각 절반은 검증 데이터로, 나머지 절반은 테스트 데이터로 활용하였다.

Mixture 음향 데이터는 수집한 데이터 셋을 거리에 따라 합성해서 생성하였으며, 사용된 음원(음성) 및 자체 잡음 음향 파일은 모두 16bit, 16kHz 샘플 레이트로 샘플링해서 사용하였다.

5.2. 학습 및 성능 분석

CSEN은 Direct Mapping 방식으로 주어진 Mixture 복소 스펙트로그램으로부터 실수 범위의 목표 음향에 대한 복소 스펙트로그램을 추정하기 때문에, 손실 함수(loss function)는 회귀 문제에서 주로 사용되는 mean squared error(MSE)를 사용하였다. CSEN의 학습은 모

두 Adam 옵티마이저를 사용해서 학습시켰으며, 학습률, beta1과 beta2는 각각 0.001, 0.9, 0.999로 설정하였다. 또한 배치(batch) 사이즈는 128로 설정하였으며, 총 500epoch까지 학습을 진행하였다.

본 연구에서는 크기 스펙트로그램의 일종인 LPS를 향상시키는 DCN[14]과 채널 간 관계를 고려하지 않고 복소 스펙트로그램을 향상시킨 CSEN without C, 그리고 weighted linear transform을 통해 채널 간 관계를 고려하면서 복소 스펙트로그램을 향상시킨 CSEN을 비교 분석하였다. 각 모델의 파라미터를 정리하면 표 1과 같다.

비교 분석에 활용된 지표로는 SDR(Signal to Distortion Ratio, 신호 대 왜곡 비율)[18]과 PESQ (Perceptual Evaluation of Speech Quality, 음성 품질의 지각 평가)[19], STOI(Short-Time Objective Intelligibility, 단시간 객관적 명료도)[20]를 사용하였다.

Table. 1 number of parameters

model	Parameter
DCN(LPS)	815,943
CSEN without C	815,943
CSEN	815,117

SDR은 실제 목표 음향과 향상된 음향 간의 왜곡 정도를 나타내는 지표로 수식 (9)과 같다.

$$SDR = 10\log_{10}\frac{\|s_{target}\|^2}{\|s_{estimate} - s_{target}\|^2} \quad (9)$$

식의 s_{target}은 실제 목표 음향을 의미하고, $s_{estimate}$는 음향 향상 모델을 통해 향상된 음향을 의미한다. 따라서 SDR은 실제 목표 음향과 향상된 음향 간의 차이가 작을수록 높은 수치를 나타내고, 반대로 클수록 0에 가까운 수치를 나타낸다. 표 2의 SDR 지표를 통한 비교 결과, 실수부와 허수부 간 관계를 고려하면서 음향을 향상시킨 CSEN이 모든 거리 구간에서 높은 성능을 보였다.

PESQ는 ITU-T에서 표준으로 제정한 대표적인 객관적 음질 평가 지표로, 전화 제조업체, 네트워크 장비 공급 업체 및 통신 사업자 등 다양한 산업 영역에서 표준 평가 지표로 사용되고 있다. PESQ는 특히 통신에 사용되는 주관적인 테스트를 모델링해서 음성 품질을 평가하기 위해 개발되었으며, 목표 신호와 추정한 신호 간에 시간 정렬 과정을 거친 후, 신호의 강도를 신호의 샘플별로 계산해서 최종적으로 1.02에서 4.56 범위의 값으로 나타난다. 표 3의 PESQ 지표를 통한 비교 결과, 모든 구간에서 복소 스펙트로그램을 향상시킨 모델들이 LPS만 고려한 모델 대비 높은 성능을 보였다. PESQ 지표에서는 평균적으로는 CSEN without C와 CSEN 모델 간 성능 차이가 크게 나지 않았으나, 음원과 멀티로터 UAV 간 거리가 1m인 구간에서는 CSEN 모델이 가장 좋은 성능을 나타내었다.

STOI는 짧은 시간 구간에서의 목표 신호와 추정된 신호 사이의 상관관계를 계산해서 음성 신호에 대한 명료도(intelligibility)를 측정하는 지표이다. 표 4의 STOI 지표를 통한 비교 결과, SDR 지표를 통한 분석 결과와 유사하게 복소 스펙트로그램을 향상시킨 모델들이 높은 성능을 보였으며, 특히 실수부와 허수부 간의 관계를 고려한 CSEN 모델이 고려하지 않은 CSEN without C 모델보다 약간 높은 성능을 나타냈다.

Table. 2 Comparison of SDR

model		DCN(LPS)	CSEN without C	CSEN
dist	SNR			
1m	-15.98	7.27	13.04	13.25
5m	-24.08	3.99	8.41	8.67
10m	-29.75	-0.31	4.09	4.33
15m	-33.21	-3.46	0.90	1.08
20m	-35.69	-6.06	-1.62	-1.47

Table. 3 Comparison of PESQ

model		DCN(LPS)	CSEN without C	CSEN
dist	SNR			
1m	-15.98	1.33	1.70	1.73
5m	-24.08	1.18	1.37	1.37
10m	-29.75	1.14	1.23	1.23
15m	-33.21	1.13	1.19	1.19
20m	-35.69	1.16	1.19	1.19

Table. 4 Comparison of STOI

model		DCN(LPS)	CSEN without C	CSEN
dist	SNR			
1m	-15.98	0.78	0.88	0.89
5m	-24.08	0.65	0.76	0.77
10m	-29.75	0.53	0.64	0.65
15m	-33.21	0.45	0.55	0.56
20m	-35.69	0.41	0.49	0.50

Ⅵ. 결 론

본 연구에서는 매우 강한 잡음 환경인 멀티로터 UAV를 통해 수집된 음향으로부터 목표 음향을 추정하여 음향을 향상시키는 방법을 연구하였다. 강한 잡음 환경에서 기존의 음향 향상 연구에서 주로 사용하는 크기 스펙트로그램이나 그 변형인 LPS만 향상시킬 때 발생하는 문제점에 대해 살펴보았으며, 시간-주파수 도메인 음향의 직교 좌표계 표현인 복소 스펙트로그램을 이용해서 음향을 향상시키는 방법을 제안하였다. 특히, 복소 스펙트로그램의 실수부와 허수부 간의 관계에 대해 분석하였으며, 실수부와 허수부 사이의 관계를 바탕으로 Mixture 복소 스펙트로그램을 향상시키는 CSEN을 제안하였다.

본 연구에서 제안한 실수부와 허수부 간 관계를 고려한 CSEN과 관계를 고려하지 않은 CSEN without C, 그리고 LPS만 향상시키는 DCN과의 비교 결과, 모든 평가지표(SDR, PESQ, STOI)에서 CSEN 모델이 가장 높은 성능을 보였다. LPS만 향상시킨 DCN 과 비교했을 때, 복소 스펙트로그램을 향상시킬 경우 음향 향상 성능이 큰 폭으로 향상되었다. 또한, 실수부와 허수부 간 관계를 추가로 고려했을 때, 모델 파라미터의 큰 증가 없이도 성능이 약간 향상됨을 확인하였다.

ACKNOWLEDGEMENT

This paper was supported by the Education and Research Promotion Program of KOREATECH in 2019.

References

[1] L. Wang, and A. Cavallaro, “Acoustic sensing from a multi-rotor drone,” *IEEE Sensors Journal*, vol. 18, no. 11, pp. 4570-4582, Apr. 2018.

[2] D. Floreano, and R. J. Wood, “Science, technology and the future of small autonomous drones,” *Nature*, vol 521, no. 7553, pp. 460-466, May. 2015.

[3] K. Daniel, S. Rohde, N. Goddemeier, and C. Wietfeld, “Cognitive agent mobility for aerial sensor networks,” *IEEE Sensors Journal*, vol. 11, no.11, pp. 2671-2682, Jun. 2011.

[4] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113-120, Apr. 1979.

[5] J. S. Lim, and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586-1604, Dec. 2005.

[6] Y. Ephraim, and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109-1121, Dec. 1984.

[7] Y. Wang, N. Arun, and W. DeLiang, “On training targets for supervised speech separation,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849-1858, Aug. 2014.

[8] J. Lee, and H. J. Kang, "A Joint Learning Algorithm for Complex-Valued TF Masks in Deep Learning-Based Single-Channel Speech Enhancement Systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1098-1108, June. 2019

[9] S. J. Park, S. M. Choi, H. J. Lee and J. B. Kim, "Spatial analysis using R based Deep Learning," *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, vol. 6, no. 4, pp. 1-8, April. 2016

[10] D. Kim, "Acquiring Real Time Traffic Information Using Deep Learning Neural Networks," *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, vol. 6, no. 5, pp. 435-444, May. 2016

[11] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks,” *IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 708-712, Apr. 2015.

[12] D. S. Williamson, Y. Wang, and D. Wang, “complex ratio masking for monaural speech separation,” *IEEE/ACM transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483-492, Dec. 2015.

[13] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7-19, Oct. 2014.

[14] Y. Li, X. Li, Y. Dong, M. Li, S. Xu and S. Xiong, “Densely Connected Network with Time-frequency Dilated Convolution for Speech Enhancement,” *IEEE International Conference on Acoustics, Speech and Signal Processing*

(ICASSP), pp. 6860-6864, May. 2019.

[15] S. W. Fu, T. Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1-6, Sep. 2017

[16] Y. J. Kim and E. K. Kim, "CNN based dual-channel sound enhancement in the MAV environment," *Journal of the Korea Institute of Information and Communication Engineering*, vol. 33, no. 12, pp. 1506-1513, Dec. 2019.

[17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *Nasa Sti/recon Technical Report N*, vol. 93, Feb. 1993.

[18] E. Vincent, R. Gribonval and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462-1469, Jun. 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," In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2001.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time – Frequency Weighted Noisy Speech," *IEEE Transactions on Audio Speech and Language Processing*, vol. 19, no. 7, pp. 2125-2136, Feb. 2011.

김은경(Eun-Gyung Kim)

1983년 2월 : 숙명여자대학교 물리학과 졸업
1986년 2월 : 중앙대학교 전자계산학과 석사
1991년 2월 : 중앙대학교 컴퓨터공학과 박사
1992년 3월~현재 : 한국기술교육대학교 컴퓨터공학부 교수
※관심분야 : 딥러닝, 빅데이터, 트리즈 등

김영진(Young-Jin Kim)

2014년 7월 : 한국기술교육대학교 컴퓨터공학부 공학사
2016년 7월 : 한국기술교육대학교 컴퓨터공학부 석사
2016년 8월~현재 : 한국기술교육대학교 컴퓨터공학부 박사과정
※관심분야 : 딥러닝, 영상처리, 음성인식 등