JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 23, No. 12: 1506~1513, Dec. 2019

MAV 환경에서의 CNN 기반 듀얼 채널 음향 향상 기법

김영진[1] · 김은경[2*]

CNN based dual-channel sound enhancement in the MAV environment

Young-Jin Kim[1] · Eun-Gyung Kim[2*]

[1]Ph.D. student, Department of Computer Science & Engineering, Graduate School, 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

최근 드론과 같은 멀티로터 UAV(Unmanned Aerial Vehicle, 무인항공기)의 산업 범위가 크게 확대됨에 따라, UAV를 활용한 데이터의 수집 및 처리, 분석에 대한 요구도 함께 증가하고 있다. 그러나 UAV를 이용해서 수집된 음향 데이터는 UAV의 모터 소음과 바람 소리 등으로 크게 손상되어, 음향 데이터의 처리 및 분석이 어렵다는 단점이 있다. 따라서 본 논문에서는 UAV에 연결된 마이크를 통해 수신된 음향 신호로부터 목표 음향 신호의 품질을 향상시킬 수 있는 방법에 대해 연구하였다. 본 논문에서는 기존의 단일 채널 음향 향상 기술 중 하나인 densely connected dilated convolutional network를 음향 신호의 채널 간 특성을 반영할 수 있도록 확장하였으며, 그 결과 SDR, PESQ, STOI과 같은 평가 지표에서 기존 연구 대비 좋은 성능을 보였다.

ABSTRACT

Recently, as the industrial scope of multi-rotor unmanned aerial vehicles(UAV) is greatly expanded, the demands for data collection, processing, and analysis using UAV are also increasing. However, the acoustic data collected by using the UAV is greatly corrupted by the UAV's motor noise and wind noise, which makes it difficult to process and analyze the acoustic data. Therefore, we have studied a method to enhance the target sound from the acoustic signal received through microphones connected to UAV. In this paper, we have extended the densely connected dilated convolutional network, one of the existing single channel acoustic enhancement technique, to consider the inter-channel characteristics of the acoustic signal. As a result, the extended model performed better than the existed model in all evaluation measures such as SDR, PESQ, and STOI.

키워드 : 2채널 음향 향상, 무인항공기, 컨볼루션 신경망, 고밀도 연결성, 확장된 컨볼루션

Keywords : Dual-Channel speech Enhancement, Unmanned Aerial Vehicle(UAV), Convolutional Neural Network(CNN), Dense Connectivity, Dilated Convolution

Received 15 August 2019, **Revised** 22 August 2019, **Accepted** 30 August 2019

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

Open Access http://doi.org/10.6109/jkiice.2019.23.12.1506 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

정보 통신 및 센싱 기술 등의 발전에 따라, 무인 항공기(Unmanned Aerial Vehicle, UAV)를 사용한 데이터의 수집 및 처리, 분석에 대한 요구가 크게 증가하고 있다. UAV는 초기에는 군사적 용도로 개발되었으나, 최근에는 민간 시장에서 재난 감시, 농업, 물류, 방송, 항공 촬영 등 다양한 산업 분야로 그 영역을 확대하고 있다[1]. UAV 중 우리가 주로 드론이라 칭하는 멀티로터(multi-rotor) UAV는 2개 이상의 로터(회전날개 또는 프로펠러)를 이용해서 이착륙과 추진, 회전하는 비행체로, 특히 4개의 로터를 사용하는 쿼드콥터는 비행 조건이 안정적이고, 비용이 저렴하여 대중적으로 널리 보급되고 있다. 마이크와 같은 센서를 부착한 멀티로터UAV는 비용이 저렴하고, 지상이나 공중의 물체에서 발생하는 음향 데이터를 수집/기록하여 음원의 위치를 추적 및 분석할 수 있으므로, 재난이나 범죄 상황에서의 인명 구조, 특정 인물이나 물체의 검색/추적 등 다양한 용도로 활용할 수 있다는 장점이 있다[2, 3, 4].

그러나 멀티로터 UAV의 회전 모터와 프로펠러에서 발생되는 강한 소음과 비행 중 발생하는 바람 소음(wind noise)은 목표 음향의 음질을 급격하게 저하시키는 문제가 있다[2]. 그림 1은 목표 음향과 잡음이 포함된 음향에 STFT(Short Time Fourier Transform)를 적용해서 시각화한 스펙트로그램으로, 멀티로터 UAV에서 수집된 음향은 대체로 그림 1의 (d)와 같이 목표 음향이 두 소음에 의해 크게 가려진 형태로 나타난다. 회전 모터와 프로펠러에서 발생되는 소음은 모터의 회전 속도와 시간에 따라 불안정적(non-stationary)으로 변화한다는 특성을 가지며, 그림1의 (b)와 같이 에너지가 주파수(frequency) 영역 전반에 걸쳐 발생되기 때문에 소음 제거가 쉽지 않다[2, 5]. 바람 소음 또한 UAV의 비행속도 및 방향에 따라 불안정적으로 변화하는 특성을 갖고 있으며, 특히 바람 소음의 에너지는 그림 1의 (c)와 같이 사람의 목소리가 주로 분포하는 낮은 주파수 영역에 집중되어, 목표 음향의 음질을 크게 저하시키는 문제를 유발한다.

따라서 멀티로터 UAV에서 목표 음향을 녹음하거나 분석하기 위해서는 자체 잡음 및 바람 소음 등으로 인해 변질된 음향으로부터 목표 음향의 질을 향상시키는 음향 향상(Speech Enhancement) 기술이 필수적이다.

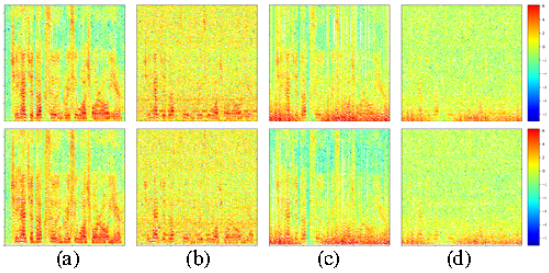

Fig. 1 real(top) and imaginary(bottom) components of spectrogram of target sound(a) and noisy sound(b,c,d)

음향 향상 기술은 시스템에서 사용되는 마이크 개수에 따라 단일 채널 음향 향상(Monaural Speech Enhancement) 기술과 다채널 음향 향상(Multichannel Speech Enhancement) 기술로 구분할 수 있다. 기존의 신호 처리 및 확률 모델 기반 단일 채널 음향 향상 기술로는 spectral subtraction [6]과 wiener filtering[7], minimum mean-squared error (MMSE)[8]를 이용한 방법 등이 있다. 그러나 이런 방법들은 잡음 및 소음이 짧은 시간 구간에서 크게 변화하지 않고 안정적(stationary)이라는 가정 하에 확률 모델을 만들기 때문에 UAV의 자체 잡음과 바람 소음처럼 매우 불안정적인 노이즈에 의해 변질된 음향의 질을 향상시키기에는 성능이 안정적이지 못하다는 단점이 있다. 최근 부상하고 있는 data-driven 방식의 Deep Neural Network (DNN) 기반 음향 향상 기술의 성능은 신호 처리 및 확률 모델 기반 음향 향상 기술의 성능을 크게 뛰어넘고 있다. DNN 기반 단일 채널 음향 향상 기술은 목표 음향의 스펙트럼을 직접 매핑하는 방식[9], Time-Frequency (T-F) mask를 추론하는 방식[10], 혹은 스펙트럼 이외의 추가적인 특징을 함께 학습하는 방식[11] 등으로 구분할 수 있다. 그중 [9]의 연구에서는 DenseNet 구조의 Dilated Convolutional Network를 사용해서 잡음이 포함된 음향의 LPS(Log Power Spectrum)로부터 목표 음향의 LPS를 효과적으로 추정하였다. 그러나 단일 채널 음향 향상 기술은 배치된 마이크의 공간적 특성을 고려하지 않기 때문에, 2개 이상의 마이크를 사용하는 환경에 적용할 경우 성능 향상이 제한적이라는 단점이 있다.

다채널 음향 향상 기술은 주로 빔 포밍(beam- forming) 기술을 적용하는데, 빔 포밍 기반 음향 향상 기술은 마이크의 개수가 많아야 성능이 보장되고, 특히 목표 음향의 방향을 알아야 적용 가능하다는 단점이 있다. 또한, 마이크 사이의 거리가 짧고 마이크의 개수가 제한될 경

우, 목표 음원의 방향을 추론하는 것이 쉽지 않다.

따라서 본 논문에서는 CNN을 이용해서 MAV(Micro Aerial Vehicle)처럼 공간적인 제약으로 인해 둘 사이의 거리가 짧은 마이크를 갖는 소형 UAV 시스템에서 취득한 음향의 품질을 향상시키는 방법에 대해 연구하였다. 본 연구에서는 2개의 마이크로부터 수신되는 신호의 특성을 모두 반영하기 위해서, densely connected dilated convolutional network[9]를 기초로 2개 마이크 채널의 스펙트럼으로부터 채널 간 특성을 반영할 수 있도록 모델의 구조를 확장했다.

Ⅱ. MAV 음향 시스템의 구성

본 연구에서는 목표 음향은 정면을 기준으로 0~180도 범위 안에서 발생한다고 가정했다. 또한, MAV의 경우 마이크를 설치할 공간이 제한되므로 그림 2의 (a)와 같이 2개의 마이크만 사용하도록 구성했다.

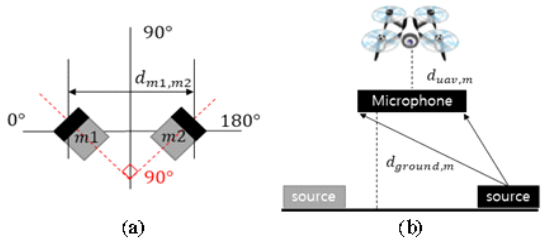

Fig. 2 Placement of Microphones

또한, 그림 2의 (b)와 같이 MAV와 2개 마이크 사이의 간격은 $d_{mav,m}$만큼 떨어져있으며, 마이크와 수집되는 음원 사이의 지면 기준 높이는 $d_{ground,m}$이다. 각 음원은 중심점 기준 5m, 10m, 20m 간격으로, 0~180° 사이를 1°간격으로 구분해서 임의로 배치했다.

마이크는 그림 2의 (a)와 같이 중심점 기준 90°의 사이 각을 갖도록 배치했으며, 마이크 사이의 거리는 $d_{m1,m2}$이다. 마이크가 전방향성(omnidirectional)인 경우 두 마이크 사이의 Inter Level Difference(ILD) 특성을 쉽게 구분할 수 없으므로, 본 연구에서는 방향성을 갖는 심장형(cardioid) 마이크를 사용했다.

Ⅲ. 제안한 CNN 모델의 구성

본 논문에서는 densely connected dilated convolutional network[9]를 기초로, 2개의 스펙트럼 간 특성을 반영할 수 있도록 확장하여 그림 3과 같이 음향 향상을 위한 CNN 모델을 구성했다. 기존 연구에서는 음향 향상에 있어서 위상(phase) 정보가 크게 중요하지 않다는 가정 하에, 예측된 LPS로부터 얻어진 스펙트럼의 크기(magnitude)와 잡음이 포함된 원본 스펙트럼의 위상을 이용해서 목표 음향 신호를 재구성했다[9]. 하지만 최신 연구에서 위상 정보가 음향 향상에 큰 영향을 미치는 것으로 밝혀졌다. 또한, 스펙트럼의 위상 정보는 값을 구하는 과정에서 -π에서 π 사이의 값으로 변환함에 따라, 전체적인 문맥 구조가 명확해지지 않기 때문에 예측이 어렵다는 단점이 있다[10]. 따라서 본 연구에서는 극 좌표계에서의 크기 및 위상으로의 표현 대신, 직교 좌표계에서의 복소 지수(complex exponential) 표현인 실수(real)부와 허수(imaginary)부를 사용하도록 모델의 입력을 구성했다.

모델의 입력 차원은 $T\times F\times 4$로, 각각 스펙트로그램의 시간 축(Time frame)의 개수, 주파수(Frequency) 축의 개수, 스펙트로그램 채널(Channel)의 개수를 의미한다. 스펙트로그램 채널은 2개 마이크 신호에 대해 STFT를 적용해서 얻은 스펙트로그램의 실수부와 허수부를 의미하며, 4개(2개 마이크로부터 수신된 스펙트로그램의 실수부와 허수부) 채널로 구성된다. 또한, 모델의 출력은 입력 시간 축의 중간에 대응되는 1개 스펙트럼($F\times 4$)을 출력으로 한다.

3.1. Extension Conv Block

CNN에서 수용 영역(receptive field)이란 하나의 컨볼루션 필터가 한 번에 볼 수 있는 영역을 의미하는데, 입력으로부터 전체적인 문맥정보를 파악하기 위해서는 수용영역이 클수록 좋다. 확장된 컨볼루션(dilated convolution)을 사용하면 모델의 파라미터를 크게 증가시키지 않고도 수용 영역을 크게 증가시킬 수 있다는 장점이 있다. 반면, 모델의 입력으로 사용하는 스펙트로그램은 그 크기가 작기 때문에, 확장된 컨볼루션을 사용해서 수용 영역의 크기를 크게 만드는 것이 제한적이다. 따라서 본 연구에서는 입력 스펙트로그램을 주파수 축으로 전치(permute)한 후, 1x1 크기의 32개 필터(filter)

를 갖는 2D 컨볼루션을 사용해서 시간 축을 확장한 다음, 다시 시간 축으로 전치한 후, 3x4 크기의 256개 필터를 갖는 2D 컨볼루션을 사용해서 주파수 축을 확장했으며, 이를 Extension Conv Block이라 칭했다.

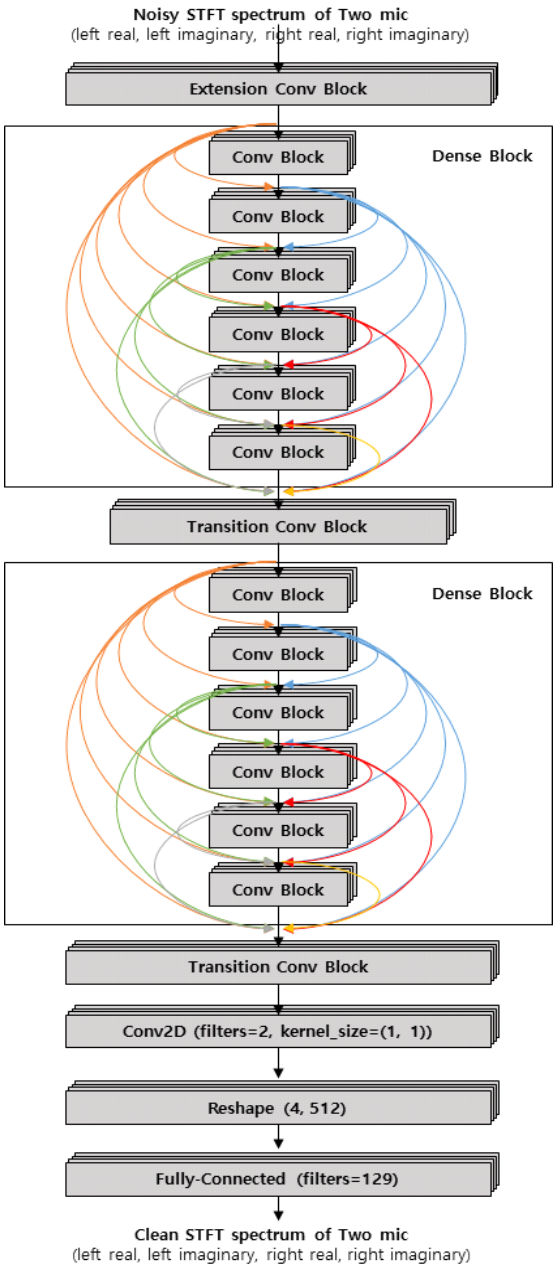

Fig. 3 CNN Architecture for sound source enhancement

3.2. Dense Block

Dense Block은 수식 (1)과 같은 dense connectivity [12]를 통해, 이전 레이어의 출력 특징 맵(feature map)을 연결되는 모든 레이어의 입력과 연결해서 구성한다. 특징 맵의 연결은 시간 축에 대해 적용했다.

$$x_l = H_l([x_0, x_1, ..., x_{l-1}]) \quad (1)$$

수식 (1)의 $[x_0, x_1, ..., x_{l-1}]$는 $0, ..., l-1$ 번째 레이어로부터 생성된 특징 맵을 연결(concatenate)한 것이며, H_l는 l번째 레이어의 비선형 활성 함수를 의미한다. dense connectivity를 사용하면 기울기 소실(vanishing-gradient) 문제가 발생하는 것을 완화할 수 있을 뿐만 아니라, 이전 레이어의 출력 특징 맵을 재사용해서 보다 효과적으로 특징을 추출할 수 있다는 장점이 있다.

Dense Block에 포함된 각각의 Conv Block의 구성은 그림 4와 같다. 시간 축의 개수 T와 주파수 축의 개수 F가 불균형하기 때문에, 각각의 시간 축과 주파수 축을 분리해서 각각 컨볼루션 연산을 취할 수 있도록 구성했다. 기존 연구와는 달리 본 논문에서는 2개 마이크로부터 수신된 음향 정보를 모두 수용할 수 있도록, 4개의 채널을 추가해서 컨볼루션 연산을 2D로 확장했다.

또한, CNN의 수용 영역을 크게 만들기 위해 확장된 컨볼루션을 사용했는데, 채널의 경우 크기가 4로 크지 않기 때문에 채널에 대한 수용 영역을 크게 만드는 것이 효과적이지 않으므로, 시간 축과 주파수 축에 대해서만 컨볼루션 연산이 확장될 수 있도록 확장 비율(dilation rate)을 (d, 1)로 설정했다.

3.3. Transition Conv Block

만약 Dense Block 내에 포함된 Conv Block의 개수가 늘어나면, 이전 레이어들에서 출력되는 모든 특징 맵이 뒤로 연결되는 모든 레이어의 입력과 연결되기 때문에 모델의 크기가 기하급수적으로 증가할 수 있다. 따라서 본 연구에서는 입력 특징의 시간 축에 대해 1x1 크기의 32개 필터를 갖는 2D 컨볼루션을 적용해서 특징 맵의 차원을 축소함으로써 모델의 파라미터가 크게 늘어나지 않도록 했으며, 이를 Transition Conv Block이라 칭했다.

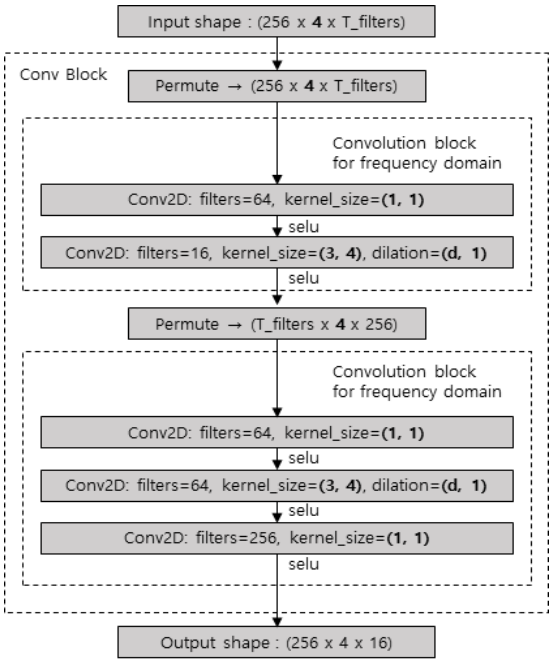

Fig. 4 The structure of Conv Block

3.4. 모델의 상세 구성

Extension Conv Block, Dense Block, Transition Conv Block을 통해 입력 스펙트로그램의 특징들이 추출되고 나면, 이후 시간 축에 대해 1x1 크기의 2개 필터를 갖는 2D 컨볼루션을 적용해서 출력 특징 맵의 차원을 간소화했다. 이후 Reshape 연산을 사용해서 특징 맵을 개별 채널 당 하나의 특징이 표현될 수 있도록 변환한 후, 각 채널에 대해 129개의 필터를 갖는 완전연결(fully-connected) 레이어를 사용해서 모델의 출력을 표현했다. 모든 레이어의 활성 함수로는 기존 연구와 동일하게 scaled exponential linear unit(SELU)[13]을 사용했으며, SELU 활성함수를 통한 self normalizing 효과를 위해 lecun_normal initializer[13]를 적용했다.

Ⅳ. 실험 및 결과 분석

4.1. 데이터 셋 구성

실험을 위한 음원에 해당하는 음성 데이터는 TIMIT 데이터 셋[14]을 사용했다. TIMIT은 630명의 화자가 발성한 총 6,300개의 음성 파일로 구성되어 있으며, 본 연구에서는 TIMIT 음성 파일들을 8:1:1 비율로 학습, 검증, 테스트 셋으로 구성했다.

MAV의 자체 잡음은 DREGON 데이터 셋[15]을 이용해서 생성했다. DREGON 데이터 셋은 MAV에 부착된 각각의 모터에 대해 속도를 달리하며 녹음한 총 21개의 자체 잡음 데이터를 포함한다. 각 잡음 데이터는 약 10~12초 길이이며, 8개의 마이크로부터 녹음된 것이다. 본 연구에서는 MAV의 자체 잡음 시뮬레이션을 위해 각각의 마이크 채널을 분리해서 총 168개의 자체 잡음 파일로 구성했으며, 각각의 잡음 음향의 시작 기준 60%인 약 6초는 학습, 그 이후 20%는 검증, 나머지 20%는 테스트 데이터 셋으로 사용했다.

바람 소음은 Multichannel Wind Noise Generator[16]를 사용해서 생성했다. 해당 Generator의 파라미터(마이크 개수, 마이크 사이 거리)는 본 연구에서 구성한 음향 시스템에 맞추어 설정했으며, 0~180도 범위에 대해 각각 30초 간격으로 총 362개의 바람 소음 데이터를 생성한 후, 각 데이터들에 대해 8:1:1 비율로 학습, 검증, 테스트 셋을 구성했다.

4.2. 시뮬레이션 및 데이터 합성

0~180도 사이에 배치되는 음원에 대한 MAV 환경의 음향 데이터를 시뮬레이션하기 위해서 gpuRIR[17]을 사용했다. gpuRIR은 Image Source Method(ISM)를 사용해서 Room Impulse Response(RIR)를 생성하는 라이브러리로, CUDA GPU를 사용한 병렬처리를 통해 빠른 시뮬레이션이 가능하다. 시뮬레이션을 위한 방의 크기(room size)는 50m, 50m, 10m 크기로 설정했으며, 일반적으로 MAV를 사용하는 환경이 야외이기 때문에 반향(reverberation)은 고려하지 않았다. 마이크 및 MAV 소음의 위치는 방의 중앙(25m, 25m)에 배치했으며, MAV와 마이크 사이의 거리 $d_{uav,m}$는 0.3m, 마이크와 음원 사이의 지면 기준 높이 $d_{ground,m}$는 1.7m로 설정했다. 또한, 마이크 간 사이 거리 $d_{m1,m2}$는 0.0395m로 짧게 설정했으며, 각 마이크의 방사 패턴은 심장형으로 설정했다.

각 모터의 소리는 DREGON[15]에서 설정한 MAV 시스템과 동일하게 중앙을 기준으로 0.2m 씩 떨어진 4개 방향에 배치하도록 설정했으며, 음원은 각각 5, 10, 20m에 대해 0~180도 사이에 1도 간격으로 배치될 수 있도록 구성했다.

생성된 모든 데이터 셋은 16kHz 샘플 레이트로 샘플

링한 후, 각 마이크 채널에 대해 256개 샘플, 50% 오버랩(overlap), 해밍(hamming) 윈도우로 설정한 Short-Time Fourier Transform(STFT)을 적용했다.

4.3. 학습 및 성능 분석

음향 향상 모델은 목표 음성의 스펙트럼을 예측하는 회귀(regression) 문제이기 때문에, 손실 함수(loss function)로 mean squared error를 적용했다. 모델의 학습은 모두 Adam 옵티마이저를 사용해서 학습시켰으며, 학습률은 0.001, beta1과 beta2는 각각 0.9와 0.999로 설정했다. 또한 배치(batch) 사이즈는 128로, 하나의 배치 당 16개의 서로 다른 음성이 포함되도록 구성했으며, epoch는 최대 500으로 설정했다.

본 연구에서는 마이크로부터 수신된 2개의 스펙트럼에 대해 기존의 단일 채널 음향 향상 기술을 적용한 모델(existed)과 채널 간 특성을 고려할 수 있도록 확장한 모델(proposed)을 SDR(Signal to Distortion Ratio, 신호 대 왜곡 비율)[2]과 PESQ (Perceptual Evaluation of Speech Quality, 음성 품질의 지각 평가)[18], 그리고 STOI(Short-Time Objective Intelligibility, 단시간 객관적 명료도)[19]라는 측정 지표에 대해 비교했다. MAV에서 수집되는 음향 데이터의 경우 대개 신호 대 잡음비(SNR)가 0dB 이하로 낮게 나타나므로, 각각 -15, -10, -5, 0dB의 SNR을 갖는 음향 데이터를 생성해서 검증에 사용했다.

SDR은 실제 목표 음성과 모델이 추정한 결과 음성 간의 왜곡 정도를 나타내는 지표로, 실제 목표 음성과 추정된 음성간의 차이가 작을수록 높은 점수로 표현된다. SDR 측정 지표에 대해 비교한 결과, 표 1과 같이 본 연구에서 제안한 모델이 모든 SNR에서 높은 성능을 보였다.

PESQ는 음성 신호의 품질을 평가할 때 주로 사용되는 지표로 －0.5에서 4.5 사이 값을 가지며, 값이 클수록 좋은 품질을 나타낸다. PESQ 지표에 대한 비교 결과(표 2), 본 논문에서 제안한 모델의 경우 표 2와 같이 SNR이 매우 낮은 -15dB에서는 성능 향상이 미미했으나, 그 이상의 SNR에서는 크게 향상된 음성 품질을 보였다.

STOI는 짧은 시간 구간에서의 목표 신호와 추정된 신호 사이의 상관관계를 계산해서 음성 신호에 대한 명료도(intelligibility)를 측정하는 지표이다. STOI 지표에 대해 비교한 결과에서도 표 3과 같이 본 연구에서 제안한 모델이 모든 SNR에서 기존 모델 대비 높은 성능을 보였으며, 특히 가장 낮은 SNR(-15dB)에서 명료도가 더 큰 폭으로 향상된 것을 확인했다.

Table. 1 Comparison of SDR

model	single channel	proposed
SNR 0	10.6424	12.1415 **(+1.4991)**
SNR -5	9.4751	10.9930 **(+1.5179)**
SNR -10	7.0171	8.3321 **(+1.315)**
SNR -15	4.6412	6.3654 **(+1.7242)**
Avg.	7.94	9.46 **(+1.52)**

Table. 2 Comparison of PESQ

model	single channel	proposed
SNR 0	1.8327	1.9710 (+0.1383)
SNR -5	1.5428	1.6572 (+0.1144)
SNR -10	1.2875	1.3423 (+0.0548)
SNR -15	1.1774	1.1889 (+0.0115)
Avg.	1.52	1.59 (+0.07)

Table. 3 Comparison of STOI

model	single channel	proposed
SNR 0	0.8927	0.9126 (+0.0199)
SNR -5	0.8372	0.8664 (+0.0292)
SNR -10	0.7487	0.7867 (+0.038)
SNR -15	0.6408	0.7108 (+0.07)
Avg.	0.78	0.82 (+0.04)

Ⅴ. 결 론

본 연구에서는 MAV에 부착된 2개의 마이크에서 수신된 음향으로부터 목표 음성의 품질을 높이는 방법에 대해 연구했다. 2개의 채널을 갖는 음향 신호로부터 목표 음성의 품질을 향상시키기 위해, 단일 채널 음향 향상을 위한 densely connected dilated convolutional network[9]를 채널 간 특성을 반영할 수 있도록 확장했다. 또한, 위상 정보가 음향향상에 영향을 미치기 때문에, 위상 정보에 대한 특성을 반영할 수 있도록 모델의 입력을 직교좌표계로 표현된 스펙트럼의 실수부와 허수부로 구성했다.

본 연구에서 제안한 모델과 기존 모델을 비교한 결과,

모든 평가 지표(SDR, PESQ, STOI)에서 본 연구에서 제안한 모델이 향상된 성능을 보였다. 단일 채널의 음향에 대한 LPS를 예측하는 기존 연구와 달리, 본 연구에서 제안한 모델이 2개 마이크에서 수신된 스펙트로그램들의 상호 간 특성을 고려함으로써 성능이 크게 향상됨을 확인했다.

본 연구에서 제안한 모델이 기존 연구보다 더 좋은 성능을 보였지만, UAV와 같이 SNR이 매우 낮은 실제 환경에 적용하기 위해서는 보다 높은 수준의 음향 향상이 필요하다. 따라서 향후 음향 향상을 보다 견고히 할 수 있는 방법에 대해 지속적으로 연구할 계획이다.

ACKNOWLEDGEMENT

This paper was supported by the Education and Research Promotion Program of KOREATECH in 2018.

References

[1] Korea Embedded Software and System Industry Association. KESSIA ISSUE REPORT [Internet]. Available: http://www.fkii.or.kr.

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

[3] 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.

[4] 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.

[5] G. Sinibaldi, and L. Marino, "Experimental analysis on the noise of propellers for small UAV," *Applied Acoustics*, vol. 74, no. 1, pp. 79-88, Jan. 2013.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] D. Wang, and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702-1726, May. 2018

[11] T. Gao, J. Du, Y. Xu, C. Liu, L. R. Dai, and C. H. Lee, "Improving Deep Neural Network Based Speech Enhancement in Low SNR Environments," In *International Conference on Latent Variable Analysis and Signal Separation*, pp. 75-82, 2015.

[12] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2261-2269, 2017.

[13] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," In *Advances in neural information processing systems*, pp. 971-980, 2017.

[14] 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.

[15] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "DREGON: Dataset and Methods for UAV-Embedded Sound Source Localization," In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1-8, 2018.

[16] D. Mirabilii, and E. A. Habets, "Simulating Multi-Channel Wind Noise Based on the Corcos Model," In *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 560-564, 2018.

[17] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for Room Impulse Response simulation with GPU acceleration," *arXiv preprint 1810.11359*, 2018.

[18] 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.

[19] 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