논문 2020-15-26

임베디드 연산을 위한 잡음에서 음성추출 U-Net 설계

(Design of Speech Enhancement U-Net for Embedded Computing)

김 현 돈*
(Hyun-Don Kim)

Abstract : In this paper, we propose wav-U-Net to improve speech enhancement in heavy noisy environments, and it has implemented three principal techniques. First, as input data, we use 128 modified Mel-scale filter banks which can reduce computational burden instead of 512 frequency bins. Mel-scale aims to mimic the non-linear human ear perception of sound by being more discriminative at lower frequencies and less discriminative at higher frequencies. Therefore, Mel-scale is the suitable feature considering both performance and computing power because our proposed network focuses on speech signals. Second, we add a simple ResNet as pre-processing that helps our proposed network make estimated speech signals clear and suppress high-frequency noises. Finally, the proposed U-Net model shows significant performance regardless of the kinds of noise. Especially, despite using a single channel, we confirmed that it can well deal with non-stationary noises whose frequency properties are dynamically changed, and it is possible to estimate speech signals from noisy speech signals even in extremely noisy environments where noises are much lauder than speech (less than SNR 0dB).

The performance on our proposed wav-U-Net was improved by about 200% on SDR and 460% on NSDR compared to the conventional Jansson's wav-U-Net. Also, it was confirmed that the processing time of out wav-U-Net with 128 modified Mel-scale filter banks was about 2.7 times faster than the common wav-U-Net with 512 frequency bins as input values.

Keywords : Speech enhancement, Noise reduction, Deep noise suppression, Deep neural network, wav-U-Net

Ⅰ. 서 론

잡음제거 기능은 실생활에서 각종 스마트 디바이스를 이용한 음성인식, 음성녹음, 전화통화, 음악감상 등에 주변의 잡음을 제거하고 음성신호를 명확하게 해주는 중요한 역할을 하고 있다. 특히 스마트폰과 AI 스피커 등에 필수적으로 음성비서 기능이 탑재됨에 따라서 음성인식 신뢰도를 높이기 위한 중요한 전처리 요소로 자리 잡았다.

고전적 잡음제거 방식은 여러개의 마이크로폰 어레이를 이용하여 음원과 노이즈의 위상차를 이용한 빔포밍 (Beamforming) 방식이 대표적이다 [1]. 하지만 마이크로폰 수와 배치 간격 그리고 알고리즘의 복잡성이 높을수록 성능이 비례하기 때문에 실용적이지 못하다. 이에 반하여 독립 성분 분석 (Independent Component Analysis, ICA) 방식은 혼합된 신호에서 독립적인 성분을 통계적으로 분리하는 원리로서 N개의 마이크로폰으로 N개의 음원 분리가 가능하다 [2]. 하지만 필터가 학습된 환경이 달라지면 적용이 어려운 단점이 있다.

실시간 임베디드 연산처리를 위한 대표적인 단채널 잡음제거 방식으로 MMSE-STSA (Minimum Mean-Square Error Short-Time Spectral Amplitude)가 있다 [3]. 보통 음성 구간 검출기 (VAD)와 연동하여 음성이 들어오기 직전의 배경잡음을 기반으로 노이즈를 제거하는 방식으로 [4], 하드웨어 성능제약이 적고 일정한 잡음 (Stationary

*Corresponding Author (reynolds@kopo.ac.kr)
Received: Jul. 28, 2020, Revised: Aug. 26, 2020, Accepted: Aug. 31, 2020.
H.-D. Kim: Korea Polytechnic (Assist. Prof.)

ⓒ IEMEK J. Embed. Sys. Appl. 2020 Oct. 15(5) 227-234
ISSN : 1975-5066
http://dx.doi.org/10.14372/IEMEK.2020.15.5.227

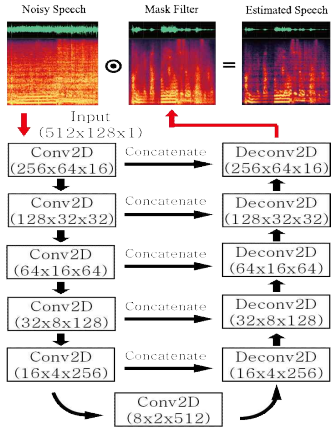

그림 1. wav-U-Net 구조

Fig. 1 Structure of wav-U-Net

noise)에는 효과적이나 음악이나 음성잡음 등과 같은 실시간 변하는 잡음 (Non-stationary noise)에는 적용이 어렵다.

최근에는 딥러닝 (Deep Learning)의 등장으로 단 채널로도 노이즈 종류에 상관없이 열악한 잡음 환경에서도 효과적으로 음성신호를 분리하는 wav-U-Net이 제안되었다 [5, 6]. 하지만 많은 층의 신경망을 가지면서 상당한 계산량의 저감 필요성과 서로 성질이 다른 잡음에 대해서 최적 성능을 내기 위한 신경망 학습이 요구된다.

이 논문에서는 임베디드 연산을 위해서 첫째로 저주파수에 주요 특징값을 보이는 음성신호를 분리를 목적으로 선형 멜 스케일 (Mel-scale) 필터뱅크를 이용한 주파수 영역 축소로 계산량을 감소시켰다 [7]. 둘째로 잡음이 섞인 음성신호를 위한 전처리 과정으로 간단한 ResNet을 거치도록 설계하였다 [8]. 따라서 wav-U-Net을 거치면서 최종 추정되는 음성신호가 명확해지고, 고주파 잡음을 억제하는 효과를 얻을 수 있다. 셋째로 잡음 특성에 따른 4가지의 DB를 사용하여 신경망을 학습하였다. 따라서 단 채널임에도 불구하고 다양한 잡음에 대응하고, SNR 0dB 이하의 극심한 노이즈 환경에서도 우수한 음성신호 분리 성능을 보여준다.

II. Wav-U-Net 배경이론

1. U-Net 알고리즘

U-Net은 의료영상 분할 (Segmentation)을 목적으로 개발되었다 [9]. U-Net은 데이터를 압축하는 경로 (인코더)와 데이터를 복원하는 경로 (디코더)가 동일한 구조로 U자 형태로 붙어있는 구조이다. 동일한 구조의 오토인코더 (Autoencoder)와의 차이점은 그림 1과 같이 서로 대응되는 convolution 층에서 인코더의 데이터가 디코더에 결합 (Concatenate)하는 것이다.

핵심 개념은 인코더에서 데이터가 압축되면서 핵심 특징값만 남게 된다. 이후 디코더를 통해서 데이터가 복원되면서 압축으로 인하여 손실된 데이터 (Missing data)는 대응되는 인코더 층 (Layer)에서 데이터를 가져와서 보정 (Extrapolation)하면서 최종 목적 신호만 복원하게 된다. 영상에서는 주로 객체검출 (Detection)과 분할 (Segmentation)에 주로 사용되며 탁월한 성능을 보여준다 [10].

2. 음악분리 wav-U-Net

wav-U-Net은 그림 1과 같이 음성신호를 대상으로 설계된 U-Net을 말한다. 인코더 과정에서 잡음 또는 특정 음원이 제거되고 디코더 과정을 통해서 원래 복원하고 싶은 음원만 남게 된다.

대표적인 wav-U-Net은 주로 음악에서 연주음과 음성을 분리하는데 적용되었다 [5, 6]. 그림 1과 같이 음성신호를 1024 크기의 FFT에서 변환된 512개의 주파수 빈 (Bin)과 128개의 프레임을 가지는 스펙트로그램 (Spectrogram) 값을 wav-U-Net에 입력하면 음성을 분리하는 마스크 (Mask)가 생성된다. 이를 원래 잡음이 섞인 음원과 요소별 합성곱 (Element-wise multiplication) 연산을 통해 최종 목적 신호인 음성만 분리된다.

III. 제안하는 wav-U-Net 구조

제안하는 wav-U-Net은 다양하고 극심한 잡음이 섞인 음원에서 음성신호를 분리할 수 있다. 여기서는 이를 구현하기 위한 네트워크 구조와 학습방법에 대해서 기술한다.

1. 제안한 멜 스케일 필터뱅크

제안한 wav-U-Net에 음성정보를 입력하기 위해서는 먼저 유효한 특징을 처리할 수 있는 형태로 변환해야 한다. 일반적으로 샘플링된 음성데이터를 FFT를 통하여 주파수 영역으로 변환시켜 사용한다. 기존의 wav-U-Net의 경우 1024 크기의 FFT에서 변환된 512 주파수 빈 (Bin) 값을 입력값으로 사용하였다 [5]. 그러나 임베디드에서는 성능 하락을 최

소화하면서 계산량 저감이 중요하므로 입력데이터 크기를 축소하는 것이 중요하다.

제안한 입력 형태는 128개의 멜 필터 뱅크 (Mel filterbank)를 사용하여 512개의 주파수 빈 개수를 128개로 축소하여 연산효율을 높였다. 여기서는 사람의 음성신호 추출이 목적이므로 멜 필터 뱅크 사용에 의한 성능 하락은 미비하다. 이유로는 음성이 성대 (Vocal cords)와 구강 (Vocal tract)으로부터 발성된다는 사실을 근거로 구강의 형태를 필터로 가정하고, 그 필터 계수를 음성의 특징으로 삼는 필터뱅크 (Filterbank)를 사용했기 때문이다. 즉, 사람이 주로 발성하는 주파수 대역은 해상도를 높이고 대부분 잡음성분이 많은 고주파 대역은 해상도를 낮추어서 음성에 대한 신호손실은 최소화하면서 고주파 대역의 잡음을 효과적으로 억제 가능한 장점이 있다. 이를 위한 멜 주파수 (Mel frequency)는 사람의 청각 특성을 반영하여 민감하게 반응하는 저주파는 해상도를 높이고 둔감한 고주파는 해상도를 낮춘 주파수로써 수식은 다음과 같다.

$$m = 2595\log_{10}(1+f/700), \quad (1)$$

$$f = 700(10^{m/2595}-1). \quad (2)$$

여기서 m은 멜 주파수를 나타내며 f는 일반적인 주파수 (Hertz)를 나타내고, 식 (1)과 (2)를 통해서 서로 변환이 가능하다.

이 논문에서는 마스크를 통해서 추출된 신호를 역 주파수변환을 통해서 신호를 복원해야 하므로 복원 시 신호의 왜곡 (Distortion)이 일어나는 삼각형 (Triangular) 필터뱅크 대신 다음 수식과 같이 정방형 (Rectangular) 필터뱅크를 사용하면서 멜 주파수 스케일만 채용하였다.

$$H_m(k) = \begin{cases} 0 & k < f(m-1),\ k > f(m+1) \\ 1 & k \geq f(m-1),\ k \leq f(m+1) \end{cases}. \quad (3)$$

여기서 H_m는 m번째 필터뱅크, k는 주파수 위치를 나타낸다.

2. 입력신호 전처리 ResNet

입력값은 제안된 멜 필터뱅크 변환을 거친 128개 행과 128프레임 열, 그리고 1차원 깊이의 3차원 형태이다. 입력값은 바로 U-Net에 제공되지 않고 사전 전처리로 그림 2와 같이 입력값을 출력값에

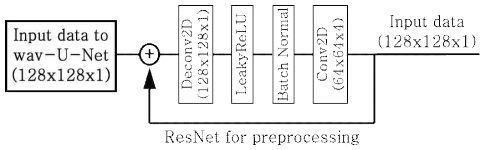

그림 2. 전처리용 ResNet

Fig. 2 Structure of pre-processing ResNet

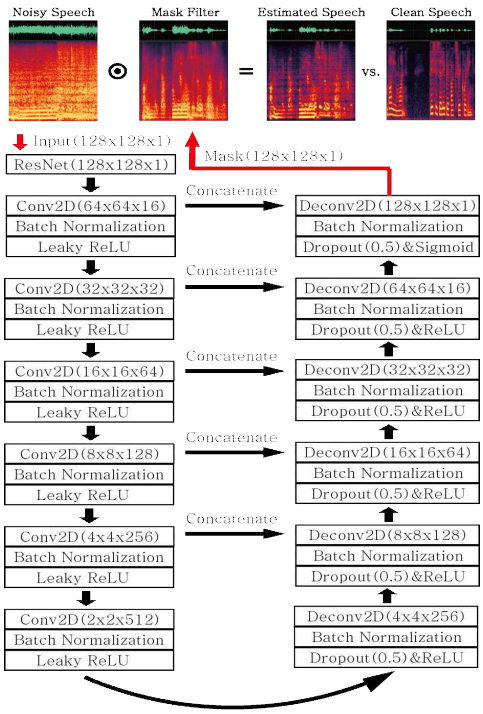

그림 3. 제안한 wav-U-Net 구조

Fig. 3 Structure of proposed wav-U-Net

더해주는 간단한 ResNet을 거친 후 입력된다. ResNet은 영상의 해상도를 높이는 데 주로 사용되는 네트워크 구조로서 일반적으로 좀 더 깊은 망을 설계할 수 있게 한다 [8]. 제안된 네트워크 구조에서는 입력신호의 고주파 잡음을 억제하고 음성신호를 명확히 해주는 효과를 볼 수 있다.

3. 제안된 wav-U-Net 구조

제안된 wav-U-Net은 그림 3과 같이 각 6개 층 (Layer)의 인코더 (Encoder)와 디코더 (Decoder)로 구성되어 있다. 인코더의 각층은 4x4 사이즈의 필터, 스트라이드 (Stride) 2, 동일한 패딩 (Same padding)의 조건을 가진 2차원 콘볼루션

(Convolution)이 있으며, 다음으로 배치 정규화 (Batch normalization) 및 Leaky ReLU 함수로 구성되어 있다.

디코더의 각층은 4x4 사이즈의 필터, 스트라이드 2, 동일한 패딩의 조건을 가진 2차원 콘볼루션 트랜스포즈 (Convolution transpose)가 있고, 배치 정규화 (Batch normalization) 이후에 50%의 드랍아웃 (Dropout)후 Leaky ReLU 함수를 통과하도록 구성되어 있다. 단, 마지막 디코더 층은 활성화 함수 (Activation function)로 시그모이드 (Sigmoid) 함수를 사용한다.

특히 U-Net의 핵심이라고 할 수 있는 데이터 결합 (Concatenate)의 경우 그림 3과 같이 첫 번째부터 다섯 번째 인코더 층을 통과한 각각의 가중치 데이터가 해당 층에 대응되는 디코더 층에 전달되어 결합이 된다. 이후 해당 디코더 층에서는 다음에 대응되는 인코더 층의 크기만큼 확장되어 진다.

이와 같은 과정을 통해서 그림 3과 같이 디코더 끝단에서는 잡음을 제거하는 마스크 (Mask)가 최종 생성이 되며 잡음이 섞인 입력신호와 요소별 합성곱 (Element-wise multiplication)을 하여 음성신호만 추출할 수 있게 된다.

4. 학습조건

제안된 wav-U-Net을 위한 데이터는 16kHz 샘플링된 음원으로서, FFT 주파수변환을 위한 윈도우 (Window) 크기는 1024을 사용한다. 한번에 입력되는 데이터 조건은 윈도우 호핑 (Hopping) 크기 128, 프레임 크기 128을 가지고 있다. 즉, 1.024초 길이단위로 처리가 된다. 그에 반해서 기존 Jansson이 제안한 음악에서 연주음과 음성을 분리하는 wav-U-Net의 경우, 윈도우 크기 800, 윈도우 호핑 크기 400, 프레임 크기 128을 가지고 있어, 3.2초 길이 단위로 처리가 된다 [5]. 따라서 본 논문에서 제안한 wav-U-Net이 훨씬 짧은 처리 시간 단위를 가지고 있으므로 임베디드 시스템에서 실시간 처리에 더 효과적이며, 세밀하게 분리 프레임 처리로 인한 손실되는 정보를 최소화 할 수 있다.

학습 (Training)에 사용되는 최적화 (Optimizer) 함수는 ADAM을, 손실 함수 (Loss function)는 MSE(Mean Squared Error)를 사용하였다. 또한, 정확한 학습의 척도를 검증 (Validation)하기 위한 검증 데이터셋의 비율은 전체 학습 데이터셋에서 20%를 할당하여 학습을 진행하였다. 최적의 학습을 위한 반복 횟수 (Epoch)는 200번 수행되었다.

Ⅳ. 성능평가

학습용 데이터셋은 음악 잡음의 경우 DS100 [11], CCmixer [12], MIR-1K [13] 데이터를 사용하였다. 그 외 잡음의 경우 Kaggle 웹사이트에서 ESC50, Urbansound8k, Cats and dogs, British birdsong, Snoring, Emergency vehicle siren, DEMAND 공개 데이터를 사용하였다. 잡음 학습을 위한 clean 음성데이터는 interspeech2020 Deep Noise Suppression Challenge를 위해 공개된 데이터를 사용하였다. 총 1,948명 화자 중 1648명을 학습 (Training)용으로 300명을 평가용 (Evaluation)으로 사용하였다 [14].

1. 학습용 데이터셋

잡음은 크게 진공청소기, 자동차 실내소음과 같이 전 주파수 영역에 분포되어 일정하게 발생하는 정체잡음 (Stationary noise)과 음악, 동물 소리와 같이 특정 주파수 영역에 분포하거나 불규칙하게 발생하는 비정체잡음 (Non-stationary noise)으로 나뉜다. 두 가지 잡음 특성은 서로 상이하므로, 일반적으로 최상의 잡음제거 성능을 얻기 위한 기존의 알고리즘 방식에도 차이가 있다.

이 논문에서는 하나의 wav-U-Net을 이용하여 다양한 잡음에 대응하여 최적의 음성신호 추출성능을 내도록 학습데이터를 구성하여 학습하였다. 먼저 정체잡음은 실내잡음 (Indoor)과 실외잡음 (Outdoor) 2개로 나누고, 비정체잡음은 음성잡음 (Vocal)과 음악잡음 (Music) 2개로 추가로 나누어서 크게 잡음 카테고리 4개로 구분하여 선별하고 학습하였다.

실내잡음은 집, 사무실, 자연환경에서 주로 발생하는 31개 정체잡음 (에어컨, 진공청소기, 세탁기, 발소리, 타이핑소리, 바람소리, 빗소리, 천둥소리 등)으로 구성되었다. 실외잡음은 식당, 카페, 길거리, 교통수단 (자동차, 지하철, 기차, 사이렌 등)에서 주로 발생하는 26개 잡음으로 구성되었다. 음성잡음은 동물 (개, 고양이, 돼지 등) 소리, 사람이 내는 잡음 (기침, 웃음소리, 코골이 등), 곤충과 새소리 등 21개 잡음으로 구성되었다. 마지막으로 음악잡음은 260곡의 연주음을 사용하였다.

4개로 분류된 잡음 데이터셋은 각각 음성 (Clean)과 잡음 (Noise)을 81,920초 분량으로 학습데이터 증가시켜 구축하였으며, 자세한 학습데이터 구성은 표 1과 같다.

표 1. 학습용 데이터셋 구성
Table 1. Contents of Training Dataset

Dataset	A	B	C	D
Category	Indoor	Outdoor	Vocal	Music
Contents (Noise)	2,889 audio clips (31 classes)	13,033 audio clips (26 classes)	6,301 audio clips (21 classes)	260 musical instruments
Contents (Clean)	Audio book clips of 1,648 people downloaded from Interspeech2020 Deep Noise Suppression Challenge			260 singing voices
Reference (Noise)	ESC50, Urbansound8k, British birdsong, Snoring, DEMAND, Cats and dogs, and Emergency vehicle siren datasets downloaded from a Kaggle website			DS100, CCmixer, MIR-1K

2. 평가방법

성능평가 지수로는 Vincent 등에 의해 개발된 SNR (Source to Noise Ratio), SDR (Source to Distortion Ratio), SIR (Source to Interferences Ratio), SAR (Source to Artifacts Ratio)의 4가지 지표를 검토하였다 [15]. 이 측정지수는 음원분리 기법 (Sound Source Separation Method)를 평가할 때 널리 사용되는 지표들이다.

상기 측정지수들에 대하여 정의를 하기 위해서, 먼저 추정된 음성신호를 다음과 같이 정의한다.

$$\widehat{S_{target}} = S_{target} + E_{noise} + E_{interf} + E_{artif}. \quad (4)$$

여기서 $\hat{S}_{target}$은 추정된 음원을 나타내며, S_{target}은 실제 음원을, E_{noise}는 실제 음원에 직접적으로 섞인 잡음을 나타낸다. E_{interf}는 간섭 잡음 (Interference noise) 신호를 말하며, 주로 녹음 시 실제 음원과 잡음 외에 주변 배경 잡음 또는 반향 신호 (Reverberation signal) 등을 나타낸다. E_{artif}는 인공 결함 잡음 (Artifact noise) 신호를 말하며, 녹음기 자체 또는 마이크로폰 자체 잡음 등으로 섞여 들어가는 에러 요소를 나타낸다. 일반적으로 시스템 잡음으로 표현하기도 하며, 성능평가 지표로 사용될 시 $E_{artif}=0$ 으로 설정하는 경우가 많다.

이제 Jansson 등이 제안한 대표적인 4가지 음원 분리기법 성능평가 지수들에 대해서 다음과 같이 나타낼 수 있으며 측정 단위는 데시벨 (dB) 이다.

먼저 일반적인 신호대 잡음의 비율을 나타내는 지표인 SNR은 다음과 같이 정의된다.

$$SNR \cong 10\log_{10}\frac{\| S_{target} + E_{interf} \|^2}{\| E_{noise} \|^2}. \quad (5)$$

실제 신호 대비 추정된 신호의 전체적인 에러에 대한 성능 지표를 나타내는 SDR은 다음과 같이 정의된다.

$$SDR \cong 10\log_{10}\frac{\| S_{target} \|^2}{\| E_{noise} + E_{interf} + E_{artif} \|^2}. \quad (6)$$

추정된 신호와 섞여 있는 간섭 (Interference) 신호의 에너지 비를 나타내는 SIR은 다음과 같이 정의된다.

$$SIR \cong 10\log_{10}\frac{\| S_{target} \|^2}{\| E_{interf} \|^2}. \quad (7)$$

추정된 신호와 시스템 잡음의 에너지 비를 나타내는 SAR은 다음과 같이 정의된다.

$$SAR \cong 10\log_{10}\frac{\| S_{target} + E_{noise} + E_{interf} \|^2}{\| E_{artif} \|^2}. \quad (8)$$

이 논문에서는 학습과 성능평가를 위해서 음원 (Clean)과 잡음 (Noise) 데이터셋을 임의로 섞어서 사용했기 때문에 $E_{interf}=0$ 및 $E_{artif}=0$ 이 되므로 SDR 지표를 사용해서 평가하였다.

또한 Jansson 등이 제안한 음악에서 연주음과 노래소리 분리용 wav-U-Net 모델을 평가할 때 사용된 Normalized SDR (NSDR)을 이용하여 추가로 평가하였으며 수식의 정의는 다음과 같다 [5].

$$NSDR(S_e, S_n, S_m) = SDR(S_e, S_n) - SDR(S_m, S_n). \quad (9)$$

여기서 S_e는 추정된 목적 신호, S_n은 순수 잡음 신호, S_m은 순수 목적 및 잡음이 섞인 신호이다.

마지막으로 성능 평가를 위한 MATLAB 코드는 공개용 BSS Eval toolkit을 사용하였다 [16].

3. 평가결과

평가용 잡음은 Interspeech2020 Deep Noise Suppression Challenge 공개 데이터을 사용하였다 [14]. 잡음 (Noise)은 65,303개 잡음 데이터셋에서 300개를 선택하였고, 음성 (Clean)은 1948명 음성 데이터셋에서 학습 (Training)용 1648화자를 제외한 300명분의 음성을 사용하였다. 최종 평가용 데이터셋은 선별된 잡음과 음성을 섞은 300개 음원을 생성하여 사용하였다.

평가는 3가지 알고리즘를 대상으로 잡음이 섞인 신호에서 음성 (Speech enhancement)을 추출하는 성능을 비교 평가하였다. 고전적인 단채널 잡음제거 알고리즘인 MMSE-STSA [3, 4], Jansson 등이 제안한 연주음과 노래소리 분리 wav-U-Net [5], 그리고 이 논문에서 제안한 다양한 잡음이 섞인 신호에서 음성신호를 추출하는 wav-U-Net에 대한 성능 비교평가를 하였다. 또한, 본 논문에서 제안한 128개의 멜 스케일 필터뱅크를 사용하지 않고 일반적인 512 주파수 빈을 입력으로 동일한 조건으로 학습한 wav-U-Net과의 성능도 비교평가 하였다. 평가 결과는 표 2와 같다.

평가결과에 따르면 기존 단채널 잡음제거 알고리즘인 MMSE-STSA과 비교하여 단채널 wav-U-Net은 압도적인 우수한 성능을 보여준다. 또한 Jansson이 제안한 wav-U-Net과 비교하면 입력 데이터로 512개의 주파수 빈 대신 단채널 128개의 멜 스케일 필터 뱅크를 사용했음에도 SDR은 약 2배 정도, NSDR은 약 5배 정도의 음성분리 성능향상이 있었다. 특히, 표 2의 하단의 결과에서 보듯이 512개 주파수 빈을 사용하는 네트워크는 제안된 wav-U-Net과 동일한 구조와 학습조건으로 생성되었지만, 128개 멜 스케일 필터 뱅크 입력값과 비교해서 4배나 큰 입력값으로 사용함에도 성능차이는 약 1% 정도로 미비함을 알 수 있다.

더욱이 300개의 평가용 데이터 셋에 대해서 전체 음원분리 실행시간을 비교하면 표 3과 같이 제안된 128개 멜 스케일 필터뱅크를 입력으로 사용한 wav-U-Net이 4배나 큰 512 주파수 빈 입력값을 사용한 것에 비해 약 2.7배 적었다. 실행조건은 라이젠 (Ryzen) 3200G, 램 16GB, SSD 저장장치에서 수행된 결과이다. 따라서 제안한 wav-U-Net이 임베디드 시스템에서 성능 하락이 거의 없으면서도 계산량이 효과적으로 감소했음을 확인하였다.

또한, 원래 음성신호와 비교하여 wav-U-Net을 통과하여 얼마나 잘 음성분리 추정이 되었는지 실내, 실외, 음성, 음악 4종류에 해당하는 대표 잡음을 선별하여 시간영역에서 음원 파평들을 비교하였다 (그림 4-7 참조). 특히, 선별된 혼합신호의 잡음의 크기는 원 음성보다 큰 (SNR 0dB 이하) 조건을 만족 하는 극심한 잡음이 섞인 음원을 분류하였다.

그림 4는 실내잡음 테스트 셋에서 진공청소기 소음이 섞여 있는 신호에서 음성 원음을 추정한 결과 파형이다. 상단부터, 혼합신호, 원 음성신호, MMSE-STSA 방식을 사용한 결과, Jansson이 제안한 wav-U-Net, 그리고 우리가 제안한 잡음제거

표 2. 잡음제거 비교평가 결과
Table 2. The results of performance

Methods / Measurement	SDR (average)	NSDR (average)
MMSE-STSA	-1.6969	-5.1148
wav-U-Net of Jansson	4.8294	1.3340
Our proposed wav-U-Net	**9.6109**	**6.1922**
wav-U-Net with 512 freq. bins	9.7108	6.2921

표 3. 평가데이터 셋 잡음제거 수행 시간
Table 3. Total execution time for evaluation

wav-U-Net input data size	128 mel-scale filter banks	512 freq. bins
Total evaluation time	10.196 sec.	27.694 sec.

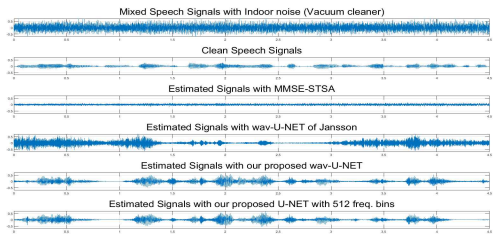

그림 4. 실내잡음 (진공청소기)에서 추정음성 파형
Fig. 4 Estimated speech plots on Indoor noise

를 위한 wav-U-Net 결과를 나타낸다. 그리고 제일 하단의 파형은 512개 주파수 빈을 입력값으로 수행했을 때의 결과이다. 그림 4와 같이 가장 상단의 혼합신호와 하단의 원음 음성을 비교해 보면 진공청소기 소음에 음성신호가 묻혀서 시각적으로 구분이 불가능하다. 이러한 극심한 잡음상황에서는 MMSE-STSA 알고리즘의 특성상 그림 4와 같이 음성분리 기능이 제대로 작동되지 못하고 실패하였다. Jansson 등이 제안한 wav-U-Net은 음악에서 노랫소리 분리에 특화된 기능을 가지고 있으므로, 진공청소기 잡음에서는 음성분리 추정이 제대로 잘 수행되지 않았음을 확인할 수 있다.

반면에 우리가 제안한 wav-U-Net에서 분리 추정된 음성신호 파형 (그림 4의 하단)은 원 음성신호 파형과 비교하여 잘 추정한 것을 확인할 수 있다. 더욱이 128개 멜 스케일 주파수와 512개 주파수 빈과 비교했을 때 시간영역에서 음원분리 추정 파형의 차이는 미미함을 알 수 있다.

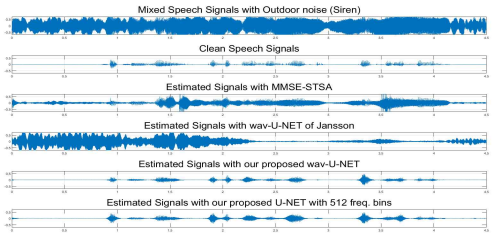

그림 5. 실외잡음 (사이렌)에서 추정음성 파형
Fig. 5 Estimated speech plots on Outdoor noise

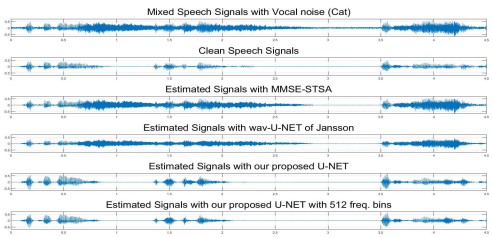

그림 6. 음성잡음 (고양이)에서 음성추정 파형
Fig. 6 Estimated speech plots on Vocal noise

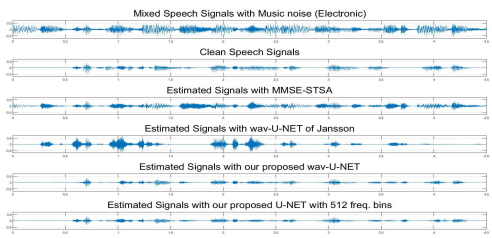

그림 7. 음악잡음 (전자음악)에서 음성추정 파형
Fig. 7 Estimated speech plots on Music noise

그림 5는 실외잡음 테스트 셋에서 사이렌 소음이 섞여 있는 신호의 음성 원음을 추정한 결과 파형이다. 그림 4와 동일하게 혼합신호에서 음성파형을 구분할 수 없는 극심한 잡음조건에서 우리가 제안한 wav-U-Net이 입력신호 크기에 상관없이 원음성신호를 잘 추정한 것을 확인할 수 있다.

그림 6는 음성잡음 테스트 셋에서 고양이 울음소리가 섞여 있는 신호의 음성 원음을 추정한 결과 파형이다. MMSE-STSA 및 Jansson이 제안한 wav-U-Net 모두 고양이와 음성 소리를 거의 구분을 못하고 혼합음과 비슷한 파형을 보여준다. 반면에 우리가 제안한 wav-U-Net의 경우 고양이 울음소리에서 음성신호 분리 추정에 독보적인 성능을 보여준다.

그림 7는 음악잡음 테스트 셋에서 전자음악 소리가 섞여 있는 신호의 음성 원음을 추정한 결과 파형이다. MMSE-STSA 경우 거의 음성분리가 되지 않은 것을 확인할 수 있다. 특히 Jansson이 제안한 wav-U-Net의 경우 음악에서 음성신호 분리에 특화된 모델이지만, 우리가 제안한 wav-U-Net이 보다 완벽하게 음악 소리를 제거하고 음성신호를 분리 추정한 것을 파형으로 확인할 수 있다.

V. 결 론

이 논문에서는 임베디드 연산을 위해서 첫째로 저주파수에 주요 특징값을 보이는 음성신호 분리를 목적으로 128개 멜 스케일 (Mel-scale) 필터뱅크를 이용한 입력데이터의 주파수 영역 축소로 계산량을 감소시켰다 [7]. 실험결과 기존 512개 주파수 빈을 입력값으로 사용했을 때보다 약 2.7배 연산 처리시간이 빨라진 것을 확인할 수 있었다. 또한, 입력 데이터 크기가 4분의 1로 축소되었음에도 실제 1% 정도의 미미한 성능 하락만 있어, 성능은 그대로 유지하면서 계산량을 줄이는 효과를 확인하였다. 둘째로 잡음이 섞인 음성신호를 위한 전처리 과정으로 간단한 ResNet을 거치도록 설계하였다 [8]. 따라서 wav-U-Net을 거치면서 추정되는 음성신호가 명확해지고 고주파 잡음을 억제하는 효과를 얻을 수 있다. 셋째로 잡음 특성에 따른 4가지의 DB (실내잡음, 실외잡음, 음성잡음, 음악잡음)를 분류하여 구축하였고, 이를 사용하여 신경망을 학습하였다. 따라서 단 채널임에도 불구하고 다양한 잡음환경에 대응이 가능할 뿐만 아니라, SNR 0dB 이하의 극심한 잡음 환경에서도 우수한 음성신호 분리 성능을 보여주었다.

비록 변형된 멜 주파수 필터 뱅크를 사용해서 입력데이터의 크기를 줄여 기존 wav-U-Net보다 계산량을 줄이긴 했지만, 여전히 네트워크 복잡성에 따른 상당한 계산량은 인공지능 기술을 임베디드에 적용하기 위해 주요 해결과제이다. 따라서 향후에는 연산량을 줄이면서도 성능을 높이기 위하여 다양한 잡음에 따라서 파라미터 가중치가 달라지는 네트워크 구조에 대해서 깊이 고찰해 볼 예정이다.

References

[1] J-M. Valin, J. Rouat, F. Michaud, “Enhanced

Robot Audition Based on Microphone Array Source Separation with Post-Filter," In IROS 2004, Sendai, Japan, pp. 2123-2128, 2004.

[2] R. Takeda, S. Yamamoto, K. Komatani, T. Ogata, and H. G. Okuno,"Missing-Feature based Speech Recognition for Two Simultaneous Speech Signals Separated by ICA with a pair of Humanoid Eras," In IROS 2006, Beijing, China, pp. 878-885, 2006.

[3] Y. Ephraim, D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," IEEE Transaction on Acoustics, Speech, and Signal Processing, Vol. ASSP-32, No. 6, pp. 1109-1121, 1984.

[4] H-D. Kim, S-S. Ahn, K. Kim, J. Choi, "Single Channel Particular Voice Activity Detection for Monitoring the Violence Situations", In 2013 IEEE RO-MAN, pp. 412-417, 2013.

[5] A Jansson, E Humphrey, N Montecchio, R Bittner, A Kumar, T Weyde, "Singing Voice Separation with Deep U-net Convolutional Networks," In ISMIR 2017, Suzhou, China, pp. 23 - 27, 2017.

[6] D. Stoller, S. Ewert, S. Dixon, "Wave-u-net: A Multi-scale Neural Network for End-to-end Audio Source Separation," In ICASSP 2018, Calgary, Canada, pp. 2391-2395, 2018.

[7] Douglas O'Shaughnessy, "Speech Communication: Human and Machine," Addison-Wesley. New York, pp. 150, 1987.

[8] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, J. Matas, "DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks", In IEEE/CVF, Salt Lake City, UT, USA, pp. 8183-8192, 2018.

[9] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", In MICCAI 2015, Springer, Vol. 9351, pp. 234-241, 2015.

[10] W. Wang, K. Yu, J. Hugonot, P. Fua, M. Salzman, "Recurrent U-Net for Resource-Constrained Segmentation", In ICCV 2019, Seoul, South Korea, pp. 2142-2151, 2019.

[11] Z. Rafii, A. Liutkus, F.-R. Stter, S.-I. Mimilakis, R. Bittner, "The MUSDB18 corpus for music separation," 2017.

[12] A. Liutkus, D. Fitzgerald, Z. Rafii, "Scalable audio separation with light kernel additive modelling," In ICASSP 2015, Brisbane, Australia, pp. 76 - 80, 2015.

[13] C.-L. Hsu, J. R. Jang, "On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset," IEEE Transactions on Audio Speech and Language Processing, Vol. 18, No. 2, pp. 310-319, 2010.

[14] C. K. A. Reddy, E. Beyrami, H. Dubey, V. Gopal, R. Cheng, R. Cutler, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Speech Quality and Testing Framework", 2020.

[15] E. Vincent, R. Gribonval, C. F´evotte, "Performance Measurement in Blind Audio Source Separation," IEEE Transactions on Audio, Speech, and Language Processing, Nol. 14, No. 4, pp. 1462 - 1469, 2006.

[16] E. Vincent, S. Araki, P. Bofill, "The 2008 Signal Separation Evaluation Campaign: A community-based Approach to Large-scale Evaluation," In ICA 2009, Paraty, Brazil, pp 734-741, 2009.

Hyun-Don Kim (김 현 돈)

He received a Ph.D. degree in Graduate School of Informatics, Kyoto University, Japan in 2008. He has been an assistant Professor at Department of Robot Automation in Robot Campus of Korea Polytechnic University.

Email: reynolds@kopo.ac.kr