Lightweight U-Net Based Monaural Speech Source Separation for Edge Computing Device

Kwang Myung Jeon  
AI Convergence Technology Laboratory  
IntFlow Co., Ltd.  
Gwangju, South Korea  
kmjeon@int-flow.com

Chanjun Chun  
Future Infrastructure Research Center  
Korea Institute of Civil Engineering and Building Technology  
Goyang, South Korea  
chanjunchun@kict.re.kr

Gyeongbong Kim  
AI Convergence Technology Laboratory  
IntFlow Co., Ltd.  
kkb8594@gmail.com

Chaejun Leem  
AI Convergence Technology Laboratory  
IntFlow Co., Ltd.  
Gwangju, South Korea  
cjleem@int-flow.com

Bohee Kim  
AI Convergence Technology Laboratory  
IntFlow Co., Ltd.  
Gwangju, South Korea  
33gksqhgml@naver.com

Wooyeol Choi*  
Department of Computer Engineering  
Chosun University  
Gwangju, South Korea  
wyc@chosun.ac.kr

Abstract—In this paper, a lightweight U-Net based monaural speech source separation method to implement high-quality speech source separation functionality in an edge computing device having a monaural microphone is proposed. The proposed method utilizes U-shaped neural networks to segregate speech and interfering noises from input mixtures in the time-frequency domain. To reduce the sizes of the networks suitable for real-time operation at the resource-constrained edge-computing device, the proposed method employs the inception-like multi-lane dimensionality reduction module for each convolutional layer of the U-Net. The performance of the proposed method is evaluated in terms of separation quality and number of parameters. Compared with the conventional U-Net based speech separation model, the proposed lightweight U-Net based method achieved a performance almost on-par with those of the conventional one while using a model footprint of 1.39 MB, which is only 3.72% of the size of the conventional U-Net. Moreover, the proposed method is successfully implemented in an off-the-shelf edge computing device having a tensor processing unit.

Keywords—speech source separation; U-Net; lightweight neural networks; edge computing

I. INTRODUCTION

In recent years, speech separation has been dramatically improved by incorporating deep learning. Various types of neural networks, such as simple multi-layer neural networks, deep recurrent neural networks, and WaveNet, have been utilized to drive the improvement of speech separation technology [1–3]. Recently, the U-Net structure, which has been successfully used for medical image separation [4], has been implemented for speech separation, and its excellent performance even in harsh environments such as wind noise and drone noise was confirmed [5], [6]. However, U-Net generally comprises more than 10 convolutional layers, and the number of network parameters ranges from 10 million to 100 million, which makes real-time execution on a typical smart device unfeasible [7].

In this paper, we propose a high-quality sound source separation method that can run in real time on an edge computing device without the help of a server or network using the Lightweight U-Net (LU-Net) model. The proposed LU-Net is essentially based on the U-Net type speech separation neural network, which separates the target speech and noise from the time-frequency domain of the audio signal. In this case, the encoder and decoder are usually composed of multi-layer 2D convolution operations. Hundreds to thousands of filters are considered for each inter-layer operation, which increases the number of parameters required for learning. Consequently, the amount of computation and the footprint of the network become very large. To overcome this problem, the proposed method significantly reduces the number of parameters required for actual learning by applying a dimensionality reduction module between 2D convolution operations. In addition, the weight of each parameter constituting the network is quantized into 8 bits to reduce the overall size and amount of computation in the network. By reducing the learning parameters of the network, the proposed LU-Net can be run in real time on resource-constrained edge computing devices, unlike conventional U-Net.

The remainder of this paper is organized as follows. Section II presents the details of the proposed lightweight U-Net based speech source separation. In Section III, we evaluate the objective speech separation performance of the proposed method and the footprint of the model. Section IV describes the details of implementing the proposed method on the edge computing device, and then Section V concludes our findings.

II. PROPOSED LIGHTWEIGHT U-Net BASED SPEECH SOURCE SEPARATION

A. Monaural Speech Source Separation

Let us consider a microphone whose input is a mixture of speech and noise signals. Subsequently, each block of the input signal, called a frame, is transformed into the frequency domain
by applying a $K$-point short-time Fourier transform (STFT). Let $Y_i(k)$ be the $k$-th spectral component at the $i$-th frame of the input signal, represented as

$$Y_i(k) = X_i(k) + D_i(k)$$

(1)

where $X_i(k)$ and $D_i(k)$ denote the $k$-th spectral component of speech and noise at the $i$-th frame, respectively. The purpose of the monaural speech source separation is to obtain $\hat{X}_i(k)$, which is an estimation of $X_i(k)$ from $Y_i(k)$. Many speech source separation methods approximate (1) into

$$|Y_i(k)| \approx |X_i(k)| + |D_i(k)|,$$

(2)

thus, the problem separates $|Y_i(k)|$ into $|\hat{X}_i(k)|$ and $|\hat{D}_i(k)|$, which are spectral magnitude estimates of speech and noise, respectively. Details regarding the separation of the spectral magnitudes using the proposed LU-Net are presented in following subsection.

**B. Lightweight U-Net Architecture**

Fig. 1 depicts the network structure of the proposed lightweight U-Net, where the numbers in each layer indicate its output shape.

$$H_i(k) = \frac{|\hat{X}_i(k)|}{|\hat{X}_i(k)| + |\hat{D}_i(k)|}$$

(5)

Finally, the clean speech estimates are obtained as

$$x_i(n) \approx \text{ISTFT}\{H_i(k)Y_i(k)\}$$

(6)

where ISTFT{∙} denotes the inverse short-time Fourier transform that reverts the spectrum into a time signal.

**III. PERFORMANCE EVALUATION**

To evaluate the performance of the proposed LU-Net, we prepared neural network models of a conventional U-Net and multiple LU-Net, which were generated from various $L_s$. Subsequently, the U-Net and LU-Net models were compared in terms of speech and noise separation performance and the size of the footprint.

**A. Experimental Setup**

First, $L_s$ for the comparison of the proposed LU-Net was set to 5, 1, 2, 4, 8, and 16; thus, 5 different LU-Net models were prepared. To train every learnable parameter for each of the six different models in this experiment, the noisy speech database for training the speech enhancement algorithms and the TTS models, released by Edinburgh DataShare [10], were used. Specifically, 11,572 clean speech utterances spoken by 28 speakers

---

**Fig. 1.** Network architecture of the proposed lightweight U-Net, where the numbers in each layer indicate its output shape.

**Fig. 2.** Network architecture of the multi-lane dimensionality reduction module. $C$ in the figure refers to the channel size in the 2D convolution.

$$\{[\hat{X}_i(k)], [\hat{D}_i(k)]\} = L_{L_s}\{Y_i(k)\}.$$  

In addition, the objective function of $L_{L_s}$ is defined by the mean-absolute error as

$$J_{MAE} = \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} \left( ||\hat{X}_i(k) - X_i(k)|| + ||\hat{D}_i(k) - D_i(k)|| \right).$$

(4)
extracted from the Voice Bank corpus [11] were prepared as the training dataset, where the average length of the utterances was measured as 2.92 s. Then, these speech utterances were added with one of 10 different noises, including two different types of artificial noises [10] and eight different types of noises taken from the diverse environments multi-channel acoustic noise database (DEMAND) [12] under different SNR conditions in the range of 0–15 dB at 5-dB steps, where the noise from DEMAND was categorized as DKITCHEN, OMEETING, PCAFETER, PRESTO, PSTATION, TCAR, TMETRO, and STRAFFIC. Thus, the training dataset comprised approximately 9 h and 23 min of noisy speech utterances, as well as the corresponding clean speech and noises.

For the evaluation, 824 clean speech utterances spoken by two speakers were also taken from the Voice Bank corpus, ensuring that none of the utterances in this evaluation dataset belonged to the training dataset. Then, each clean utterance was mixed with five different types of noise, namely SCAFE, OOFFICE, TBUS, DLIVING, and SPSQUARE, from the DEMAND database, under four different SNR conditions, namely 17.5, 12.5, 7.5, and 2.5 dB.

B. Quality versus Footprint Comparison

To compare the performance of U-Net and the proposed LU-Net with different $L_s$, we compared the MAEs of the results of processing the evaluation dataset with the footprint of each network.

Fig. 3 depicts the network footprint compared to the average MAE of speech and noise yielded from each source separation model.

Next, Fig. 4 shows the spectrogram comparison of the speech sources used in the evaluation. Fig. 4(a) displays the spectrogram of the noisy speech signal that was prepared by mixing clean speech signal with bus noise to have 2.5 dB on average. Next, Figs. 4(b) and 4(c) show the source separated result yielded by the conventional U-Net, which could be considered as the desired anchor. Subsequently, the source separation results obtained from the proposed LU-Net with $L_s = 1$ yielded results that were indistinguishable from those of U-Net in terms of a subjective comparison, as shown in Figs. 4(d) and 4(e). Finally, Figs. 4(d) and 4(e) indicate that the proposed LU-Net with $L_s = 16$ separated the speech and noise that followed the spectral distribution, which was highly similar to those of U-Net, and LU-Net with $L_s = 1$, with negligible difference in terms of ambiance level. From these results, the proposed LU-Net could separate speech and noise as successfully as the previous U-Net with a substantially smaller network footprint.

IV. IMPLEMENTATION INTO EDGE COMPUTING DEVICE

To implement the proposed LU-Net-based monaural source separation method for real-time audio recording applications, a resource constrained single-board computer having a single-channel microphone was prepared as the target device, and then further optimization of the proposed neural networks was conducted. First, the target device was equipped with a 32bit quad-core ARM processor that operates Debian Linux. A single MEMS microphone sensor embedded in the device was used to record audio signals with a sampling rate of 16 kHz and a 16bit resolution. Next, the proposed LU-Net with $L_s = 16$ was prepared and then quantized to have a data type of unsigned integer (8bit), instead of float (32bit); thus, the size of the quantized ver-
This paper proposed LU-Net, a lightweight version of U-Net for monaural speech source separation in the time-frequency domain. The proposed LU-Net had an MLDR module with a hyperparameter, $L_s$, to control the trade-off between the separation performance and the footprint. An objective evaluation indicated that the proposed method had been comparable with $L_s = 1$ and yielded a comparable performance to the U-Net in terms of MAE while achieving a model footprint of 1.39 MB, which was only 3.72% of the size of the conventional U-Net. The proposed method was successfully implemented on the resource constrained single board computer, and we plan to demonstrate it during our presentation at the conference.

REFERENCES