TEACHER-STUDENT LEARNING FOR LOW-LATENCY ONLINE SPEECH ENHANCEMENT USING WAVE-U-NET

Sotaro Nakaoka, Li Li, Shota Inoue, Shoji Makino

University of Tsukuba, Japan

ABSTRACT

In this paper, we propose a low-latency online extension of wave-U-net for single-channel speech enhancement, which utilizes teacher-student learning to reduce the system latency while keeping the enhancement performance high. Wave-U-net is a recently proposed end-to-end source separation method, which achieved remarkable performance in singing voice separation and speech enhancement tasks. Since the enhancement is performed in the time domain, wave-U-net can efficiently model phase information and address the domain transformation limitation, where the time-frequency domain is normally adopted. In this paper, we apply wave-U-net to face-to-face applications such as hearing aids and in-car communication systems, where a strictly low-latency of less than 10 ms is required. To this end, we investigate online versions of wave-U-net and propose the use of teacher-student learning to prevent the performance degradation caused by the reduction in input segment length such that the system delay in a CPU is less than 10 ms. The experimental results revealed that the proposed model could perform in real-time with low-latency and high performance, achieving a signal-to-distortion ratio improvement of about 8.73 dB.

Index Terms— Wave-U-net, single-channel speech enhancement, teacher-student learning, low-latency, real-time

1. INTRODUCTION

Since noise inevitably reduces the intelligibility and quality of speech in real-world environments, speech enhancement techniques [1] are used in various speech processing systems, such as speech recognition systems, hearing aid devices, teleconference systems, and in-car communication.

Recent advances in deep neural networks (DNNs) have dramatically improved the performance of monaural speech enhancement [2]. A wide variety of network architectures have provided various approaches [3, 4, 5, 6] to accomplish speech enhancement in the time-frequency (TF) domain with high performance. The general idea of these methods is to train a DNN to learn a nonlinear mapping from spectral magnitudes of a noisy speech obtained with short-time Fourier transform (STFT) to those of a clean speech or a TF mask. The waveform of an enhanced speech is then obtained by applying inverse STFT (iSTFT) using the enhanced magnitude and noisy phase. However, there are two downsides in these methods. First, the use of noisy phase limits enhancement performance. The phase information has shown to be essential for improving speech intelligibility and quality [7, 8], which should also be considered in the optimization. Although some attempts [9, 10, 11, 12] have been made to address this problem by applying phase-aware estimation and shown to boost the performance, performance limitations remain owing to the lack of constraints on STFT consistency and mixture consistency [13]. Secondly, effective source separation in the frequency domain requires high frequency resolution, which is obtained over a long analysis window. This results in a relatively high system latency in real-time applications since the window length bounds the minimum latency.

Another promising way to address these problems is by directly performing source separation in the time domain [14, 15, 16]. Wave-U-net is one such method, originally proposed for singing voice separation [16] and then applied to speech enhancement [17]. Wave-U-net uses a one-dimensional (1D) convolutional neural network (CNN) with a series of downsampling and upsampling blocks to estimate a clean speech when an utterance with noise is input. Since wave-U-net does not perform STFT, there is no need to consider STFT consistency or high frequency resolution. Moreover, the estimated noise signal is obtained by suppressing the estimated speech signal from the mixtures so that the mixture consistency holds. Attracted by these useful properties and high performance of wave-U-net, in this paper, we extend wave-U-net to an online method, making it suitable for face-to-face applications, where latency of less than 10 ms is a prerequisite [18]. Although this can be achieved by simply applying block processing using the original wave-U-net model owing to the fully convolutional network (FCN) architecture, the enhancement performance degrades since the clues from input segments available for inference are reduced. To prevent degradation, we propose using teacher-student learning (also known as knowledge distillation) [19] to train the model that can estimate a clean speech with short input segments.

Related work: In [20], a model similar to wave-U-net that uses CNNs and encoder-decoder architecture with skip connections has been proposed. Although the method can

This work was supported by JSPS KAKENHI Grant Number 19H04131 and Strategic Core Technology Advancement Program (Supporting Industry Program).

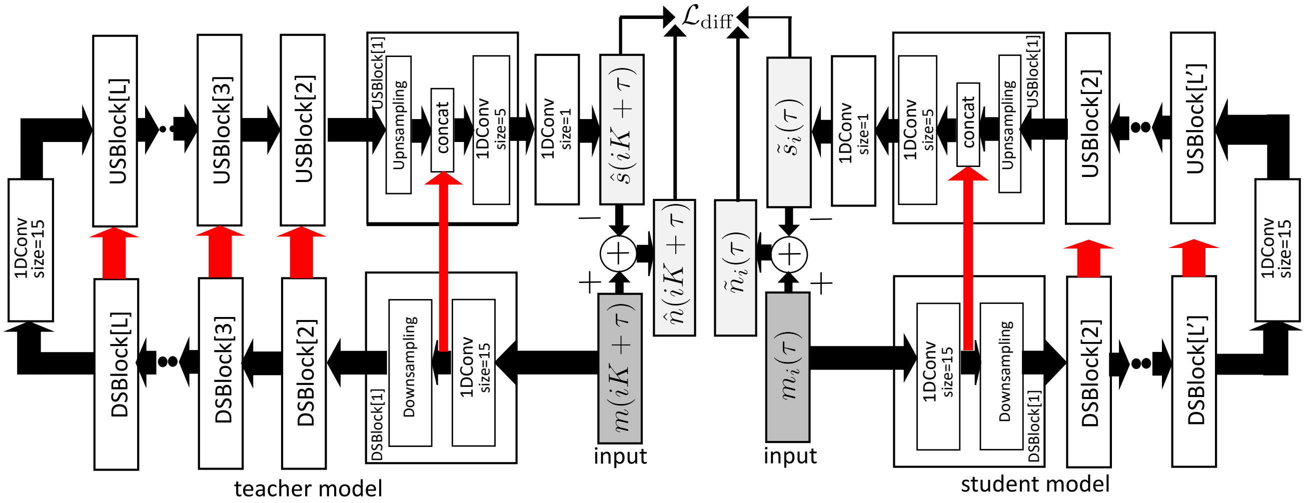

Fig. 1. Architectures of offline and online wave-U-net, which correspond to teacher and student models.

work in real-time, it is challenging to reduce window length and stride further to meet the latency requirement since the processing time was longer than 10 ms. In [21], a low-latency model using a temporal convolutional neural network (TCNN) is proposed, where an additional module consisting of causal and dilated convolutional layers was utilized to capture longer-time dependencies of encoded features. Compared with this model, the proposed model is more compact, where the parameter number is halved.

The rest of the paper is organized as follows. In Section 2, we review details of wave-U-net for monaural speech enhancement. In Section 3, we first describe blockwise processing for the online algorithm and then introduce the training strategy using teacher-student learning. In Section 4, we provide experimental evaluation settings and results, and in Section 5,we conclude this paper.

2. WAVE-U-NET FOR SPEECH ENHANCEMENT

Given an input observed mixture signal $m(t) = s(t) + n(t)$, where $m(t) \in [-1, 1]$, $s(t) \in [-1, 1]$, and $n(t) \in [-1, 1]$ are the mixture signal, target speech signal, and noise signal, respectively, the aim of wave-U-net is to estimate the clean speech signal

$$\hat{s}(1), \ldots, \hat{s}(T) = \mathcal{F}_\theta(m(1), \ldots, m(T)). \tag{1}$$

Here, $\mathcal{F}_\theta(\cdot)$ represented by a neural network with parameter θ denotes a nonlinear mapping from the mixture signals to the clean speech signal and $t = [1, \ldots, T]$ denotes the time index. To meet the mixture consistency, the estimated noise signal is computed as

$$\hat{n}(t) = m(t) - \hat{s}(t). \tag{2}$$

The network parameters θ are trained by minimizing the mean square error (MSE) between the estimated and clean signals, which is expressed as

$$\mathcal{L} = \mathbb{E}\Big[|\hat{s}(t) - s(t)|^2 + |\hat{n}(t) - n(t)|^2\Big]. \tag{3}$$

For the network architecture, wave-U-net is designed to consist of an encoder and a decoder, which are composed of L downsampling (DS) blocks and L upsampling (US) blocks followed by a 1D convolutional layer each, namely, the bottleneck layer and output layer. An illustration of the wave-U-net architecture is shown as the teacher model in Fig. 1. The DS blocks stack a 1D CNN layer with a nonlinear function and a downsampling layer that halves the feature resolution by discarding every other feature. The US blocks stack an upsampling layer and a 1D CNN layer with nonlinearity, where the upsampling layer applies convolution after recovering intermediate values by interpolation to double the feature resolution. Note that instead of using transposed convolution layers that apply convolution after padding zeros between original values, performing upsampling with interpolation can prevent aliasing artifacts caused by zeros, which may degrade enhancement performance. Except the output layer where tanh nonlinearity is used to constrain the output values in the interval of $[-1, 1]$, the nonlinear functions used in the network are Leaky ReLU [22].

3. PROPOSED ONLINE LOW-LATENCY MODEL

Wave-U-net has shown to achieve impressive performance in speech enhancement tasks [17]. With the motivation of maintaining the high performance of wave-U-net while applying it to face-to-face applications, in which a system latency longer than 10 ms is unacceptable for users, we propose an online wave-U-net model with a compact size. To this end, we apply teacher-student learning for knowledge transfer. In this section, we first describe the online algorithm with blockwise processing. We then introduce teacher-student learning and explain how it is applied to the wave-U-net.

3.1. Online wave-U-net

One simple way to apply real-time processing is to perform enhancement for each input segment $m_i(\tau) = m(iK + \tau)$, where $\tau = [1, \dots, K]$ denotes the time index in the ith segment and K is the segment length. As the lower boundary of system latency is K, we would like to reduce the segment length. However, this may reduce the enhancement performance since less information is available for inference. The tradeoff between performance and segment size needs to be considered for applications that strictly require low-latency processing.

Since future information is not available for online algorithms, causal CNN are regularly used in real-time processing. Moreover, compared with applying regular CNNs to signals, it is expected that causal CNNs will generate fewer artifacts in segmental boundaries because zero-padding is performed on one side only. An experimental comparison between using regular and causal CNNs will be given in Section 4.

3.2. Teacher-student learning for knowledge transfer

Teacher-student learning [19] is a method of knowledge transfer using a well-pretrained teacher network to teach a student network to make the same inference as the teacher network does. It is usually used for domain adaptation [23, 24] and transfering the information between networks with different architectures, such as from deep to shallow [25] or different types of networks [26]. With a pretrained teacher network that outputs estimated signals, the student network is usually trained by minimizing the following loss function:

$$\mathcal{L}_{\text{stu}} = \mathcal{L} + \beta \mathcal{L}_{\text{diff}}, \tag{4}$$

where $\mathcal{L}$ is a loss function measuring the estimation accuracy using clean signals and $\mathcal{L}_{\text{diff}}$ measures the dissimilarity between outputs of the teacher and student models. Here, $\beta \geq 0$ is a parameter that weighs the importance of both terms.

As mentioned above, reducing the input segment length is necessary to reduce system latency. Therefore, we would like to train a student network $\mathcal{G}_\phi(\cdot)$ to accurately infer the target speech $\tilde{s}_i(\tau)$ with a short input segment $m_i(\tau)$, i.e., $\tilde{s}_i(1), \dots, \tilde{s}_i(K) = \mathcal{G}_\phi(m_i(1), \dots, m_i(K))$. To guide the training of this student model, we can use the offline wave-U-net as a teacher model. The parameter ϕ is trained using (4), where the first term is defined as (3) and the second term is defined using the output of the offline wave-U-net as

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\Big[|\tilde{s}_i(\tau) - \hat{s}(iK + \tau)|^2 + |\tilde{n}_i(\tau) - \hat{n}(iK + \tau)|^2\Big]. \tag{5}$$

Here, $\tilde{n}_i(\tau)$ is the estimated noise signal computed by suppressing the output of the student model $\tilde{s}_i(\tau)$ from the observed mixture signal $m_i(\tau)$. Fig. 1 illustrates the flowchart of teacher-student learning. The network architecture for the student model is designed to have a structure similar to that of the offline wave-U-net, which consists of L' DS and US blocks.

Table 1. Parameter number of different student models.

model	parameter #
$L' = 6$	1,079,302
$L' = 7$	1,625,602
$L' = 8$	2,329,942

Table 2. Average SDR, SIR, and SAR [dB] achieved by models using regular CNN and causal CNN with $K = 1024$.

model	SDR	SIR	SAR
regular CNN	7.82	14.31	9.55
causal CNN	7.51	13.15	9.38

4. EXPERIMENTS

We conducted several experiments to evaluate the speech enhancement performance of offline wave-U-net, online wave-U-net with block processing, and models trained with teacher-student learning for an in-car communication system. Signal-to-distortion ratios (SDRs), source-to-interferences ratios (SIRs), and sources-to-artifacts ratios (SARs) [27] were used to evaluate the enhancement performance, and perceptual evaluation of speech quality (PESQ) [28] and short-time objective intelligibility (STOI) [29] were used to evaluate the speech quality and intelligibility.

4.1. Datasets

We excerpted utterances of clean speech spoken by 10 speakers from the CMU Arctic database [30], including 100 utterances for each speaker. 6 speakers (4 males and 2 females) labeled as {“aew”, “ahw”, “aup”, “axb”, “eey”, “fem”} were used to generate the training and validation datasets, and the other speakers (2 males and 2 females) labeled as {“awb”,“bdl”, “clb”, “slt”} were used for the test. All utterances were about 3 to 7 seconds long. We excerpted noise signals from the JEIDA-NOISE database [31], which were recorded in two different types of cars that moved at a high speed. There was about 1-hour data available for each type. We used noise recorded in one car as training data and the other one for the test. The mixture signals for training and validation were generated by adding the speech utterances to randomly segmented noise signals, whose power was controlled by multiplying a randomly selected scaling parameter in [0.2, 0.9], whereas the mixture signals for the test were generated by setting signal-to-noise ratios (SNRs) at {-3, 0, 3} dB. For each SNR, 50 test samples were generated. The

Table 3. Average SDR, SIR, SAR, PESQ, and STOI achieved by models trained without or with teacher-student learning, where $\beta = 0$ and $\beta = 0.01$ denote without and with teacher-student learning, respectively.

model	K	SDR [dB]		SIR [dB]		SAR [dB]		PESQ		STOI	
		$\beta = 0$	$\beta = 0.01$	$\beta = 0$	$\beta = 0.01$	$\beta = 0$	$\beta = 0.01$	$\beta = 0$	$\beta = 0.01$	$\beta = 0$	$\beta = 0.01$
online	64	4.79	8.80	11.62	20.36	6.34	9.34	2.14	2.34	0.85	0.87
	128	6.07	10.35	13.85	24.12	7.22	10.66	2.40	2.58	0.88	0.89
	256	7.83	11.29	14.68	25.63	9.17	11.55	2.54	2.73	0.91	0.91
	512	7.08	11.80	14.24	27.43	8.55	11.99	2.70	2.80	0.92	0.92
	1024	7.82	12.19	14.31	28.07	9.55	12.38	2.72	2.90	0.93	0.93
	2048	8.28	12.22	16.38	28.54	9.68	12.38	2.78	2.94	0.92	0.93
offline		13.84	—	29.81	—	13.98	—	2.56	—	0.92	—

Table 4. Average processing time and system latency [ms] with different input segment sizes.

K	processing time	system latency
64	2.71	6.71
128	3.40	11.40
256	4.83	20.83
512	6.05	38.05
1024	8.62	72.62
2048	14.89	142.89

average SDR of input mixture signals was 0.07 dB. All the signals were sampled at 16 kHz.

4.2. Network architectures and training settings

The offline wave-U-net model consisted of 8 DS and US blocks, i.e., $L = 8$. We set the convolutional filter size for DS blocks, US blocks, the bottleneck layer, and the output layer at 15, 5, 15, and 1, respectively. The channel number for the lth DS block and US block was $20l$. For the bottleneck layer and output layer, the channel number was set at $20(L + 1)$ and 1, respectively. We modified the length of training data to 64000 samples by zero-padding since the length of 64000 samples was larger than most of the utterances in the datasets. The Adam optimizer with a learning rate of 0.0001 was used for training. The minibatch size was 32.

The trained offline model was used as the teacher model to train student models, where the input segment size $K = \{64, 128, 256, 512, 1024, 2048\}$. Except for the model with $K = 64$ and $K = 128$ where the number of DS and US blocks L' was 6 and 7, respectively, other student models had the same structure as the offline model. The parameter numbers of different student models are shown in Table 1. The weight parameter β for teacher-student learning was set at 0.01.

4.3. Experimental Results

We first conducted a preliminary experiment to compare the effect of regular CNNs and causal CNNs with those of models trained for $K = 1024$. Note that we did not apply teacher-student learning in this experiment. Table 2 shows the results. We found that causal CNNs performed slightly worse than regular CNNs, which was opposite to what we expected. On the basis of these results, we conducted all the subsequent experiments using regular CNNs.

Table 3 shows the average SDR, SIR, SAR, PESQ, and STOI over 150 test samples. By comparing the results of the offline model with those of the online model without teacher-student learning, we found that the speech enhancement performance decreased significantly with shorter input segments. In particular, at $K = 64$, the SDR score decreased by more than 9 dB, which was also attributed to the smaller model. Results of models trained using teacher-student learning achieved better scores in terms of all the criteria. Specifically, the SDR scores increased by about 4 dB for all models. Although there was still some room for improvement of 5 dB in the offline model, these results demonstrated the effectiveness of using teacher-student learning.

In Table 4, we show the system latency and average time to process a single segment for each model. The entire processing was performed on an Intel (R) Core (TM) i7-7800X CPU@3.50GHz. All the models were able to work in real-time. The lowest system latency was about 6.7 ms, which was less than 10 ms. This indicates that the proposed model with its compact size can be used for face-to-face applications. The $K = 128$ model, which achieved an SDR score of about 1.5 dB higher than $K = 64$ and operated with lower latency, also showed high potential for face-to-face applications.

5. CONCLUSIONS

In this paper, an online extension of wave-U-net was proposed for face-to-face applications, where system latency is required to be less than 10 ms. Specifically, we utilized teacher-student learning for training a model that enhances noisy speech signals when short segments are input, which successfully reduced the system latency to about 6.7 ms. We conducted experiments to evaluate the speech enhancement performance by simulating a situation in a vehicle cabin. The results showed that the model trained using the proposed method performed well in real-time and with low latency.

6. REFERENCES

[1] P. C. Loizou, “Speech enhancement: Theory and practice,” *CRC press*, 2013.

[2] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Trans. ASLP*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Trans. ASLP*, vol. 23, no. 1, pp. 7–19, 2014.

[4] K. Tan and D. Wang, “A convolutional recurrent neural network for real-time speech enhancement,” in *Proc. Interspeech*, pp. 3229–3233, 2018.

[5] A. Narayanan and D. Wang, “Ideal ratio mask estimation using deep neural networks,” in *Proc. ICASSP*, pp. 7092–7096, 2013.

[6] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, “Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR,” in *Latent Variable Analysis and Signal Separation*, 2015.

[7] M. S. E. Langarani, H. Veisi, and H. Sameti, “The effect of phase information in speech enhancement and speech recognition,” in *Proc. ISSPA*, pp. 1446–1447, 2012.

[8] K. Paliwal, K. Wójcicki, and B. Shannon, “The importance of phase in speech enhancement,” *speech communication*, vol. 53, pp. 465-494, 2011.

[9] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks,” in *Proc. ICASSP*, pp. 708–712, 2015.

[10] D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ACM Trans. ASLP*, vol. 24, pp. 483-492, 2016.

[11] N. Takahashi, P. Agrawal, N. Goswami, and Y. Mitsufuji, “PhaseNet: Discretized Phase Modeling with Deep Neural Networks for Audio Source Separation,” *in Proc. Interspeech*, pp. 2713–2717, 2018.

[12] J. Le Roux, G. Wichern, S. Watanabe, A. Sarroff, and J. R. Hershey, “Phasebook and friends: Leveraging discrete representations for source separation,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 370–382, 2019.

[13] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, and R. A. Saurous, “Differentiable consistency constraints for improved deep speech enhancement” in *Proc. ICASSP*, pp. 900–904, 2019.

[14] S. Pascual, A. Bonafonte, and J. Serrà, “SEGAN: Speech enhancement generative adversarial network,” in *Proc. Interspeech*, pp. 3642–3646, 2017.

[15] Y. Luo and N. Mesgarani, “Tasnet: Time-domain audio separation network for real-time, single-channel speech separation,” in *Proc. ICASSP*, pp. 696–700, 2018.

[16] D. Stoller, S. Ewert, and S. Dixon, “WAVE-U-NET: A multi-scale neural network for end-to-end audio source separation,” in *Proc. ISMIR*, 2018.

[17] C. Macartney and T. Weyde, “Improved speech enhancement with the wave-u-net,” *arXiv preprint:1811.11307*, 2018.

[18] J. Agnew and J. M. Thornton, “Just noticeable and objectionable group delays in digital hearing aids,” *Journal of the American Academy of Audiology*, vol. 11, no. 6, pp. 330–336, 2000.

[19] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in *Proc. NIPS*, 2014.

[20] A. Defossez, G. Synnaeve, and Y. Adi, “Real time speech enhancement in the waveform domain,” *arXiv preprint:2006.12847*, 2020.

[21] A. Pandey and D. Wang, “TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain,” in *Proc. ICASSP*, pp. 6875–6879, 2019.

[22] A. L. Maas, A. Y. Hannun, and A. Y. Ng, “Rectifier nonlinearities improve neural network acoustic models,” in *Proc. ICML*, vol. 30, no. 1, 2013.

[23] Z. Meng, J. Li, Y. Zhao, and Y. Gong, “Conditional teacher-student learning,” in *Proc. ICASSP*, pp. 6445–6449, 2019.

[24] V. Manohar, P. Ghahremani, D. Povey, and S. Khudanpur, “A teacher-student learning approach for unsupervised domain adaptation of sequence-trained ASR models,” in *Proc. SLT*, pp. 250–257, 2018.

[25] A. Polino, R. Pascanu, and D. Alistarh, “Model compression via distillation and quantization,” in *Proc. ICLR*, 2018.

[26] K. J. Geras, A.-R. Mohamed, R. Caruana, G. Urban, S. Wang, O. Aslan, M. Philipose, M. Richardson, and C. Sutton, “Blending LSTMs into CNNs,” in *Proc. ICLR*, 2016.

[27] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[28] 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 *Proc. ICASSP*, Cat. No. 01CH37221, vol. , pp. 749–752, 2001.

[29] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *Proc. ICASSP*, pp. 4214–4217, 2010.

[30] J. Kominek and A. W. Black, “The CMU Arctic speech databases,” in *Proc. SSW*, pp. 223–224, 2004

[31] JEIDA Noise Database, accessed in Oct. 21, 2020, http://research.nii.ac.jp/src/en/JEIDA-NOISE.html