Reducing algorithmic delay using low-overlap window for online Wave-U-Net

Sotaro Nakaoka*, Li Li†, Shoji Makino*‡ and Takeshi Yamada*
* University of Tsukuba, Japan
† NTT Communication Science Laboratories, NTT Corporation, Japan
‡ Waseda University, Japan

***Abstract*—Wave-U-Net is an end-to-end single-channel source separation method that works in the time domain and thus can take the phase information into account during separation. It has shown high performance in tasks such as singing voice separation and speech enhancement. We previously proposed an extension of Wave-U-Net to online processing with a short input using teacher-student learning. Since online Wave-U-Net processes input signals frame-by-frame, where the frames are segmented by applying a window function, the window length is generally the lower bound of the algorithmic delay. In this paper, based on the fact that the separation performance of online Wave-U-Net is concentrated at the center of the segment, we propose to reduce the algorithmic delay by applying windows with a zero region near the edges into the online Wave-U-Net. Experimental results showed that the proposed method reduced the algorithmic delay by 40% of that of the conventional method while keeping the high speech enhancement performance with source-to-distortion ratio improvement of about 15 dB, thus enabling low-delay and high-performance speech enhancement.**

I. 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 markedly 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 the noisy speech obtained with the short-time Fourier transform (STFT) to those of clean speech or a TF mask. The waveform of enhanced speech is then obtained by applying the inverse STFT (iSTFT) using the enhanced magnitude and noisy phase. However, there are two downsides in these methods. First, the use of a noisy phase limits the enhancement performance. The phase information has been 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 have been 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 relatively high system latency in real-time applications since the window length bounds the minimum latency.

Another promising way to address these problems is to directly perform source separation in the time domain [14], [15], [16]. Wave-U-Net is one such method, which was proposed for singing voice separation [16] and then applied to speech enhancement [17]. Wave-U-Net for speech enhancement uses a one-dimensional (1D) convolutional neural network (CNN) with a series of downsampling and upsampling blocks to estimate clean speech when an utterance with noise is input. Since Wave-U-Net does not perform an 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. These characteristics are particularly suitable for conversational applications that require online processing with low latency.

On the basis of these properties, we have extended Wave-U-Net to online methods [18] and confirmed that speech enhancement with high-quality can be achieved even for block processing with short segments as input by using teacher-student learning (also known as knowledge distillation) [19]. With this extension, we have shown that online Wave-U-Net can achieve source-to-distortion ratio improvement (SDRi) of 8.73 dB with an algorithmic delay of 4 ms and SDR of 12.12 dB with an algorithmic delay of 64 ms. These results indicated that there was a tradeoff between the performance and the algorithmic delay, which is generally lower bounded by the length of input segments.

In this paper, to achieve low latency with high speech enhancement performance, we propose online Wave-U-Net with a low-overlap window to reduce the algorithmic delay instead of shortening the length of input segments. The low-overlap window has been used in speech compression [20], which has zero regions at both ends of the window and can segment signals shorter than the length of window. Therefore it is able to reduce the algorithmic delay while keeping the same input length. In addition, since Wave-U-Net directly performs enhancement in the time-domain with a network, we can expect that the effects of the frequency response of the window function can be modeled by the network and learned from the training data. Therefore, it will not have a significant

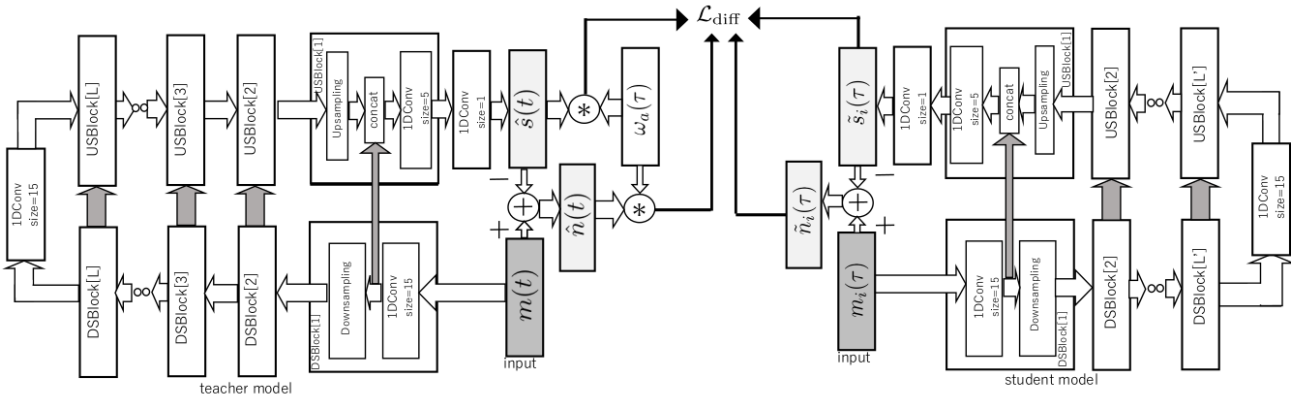

Fig. 1. Network structures of offline (left) and online Wave-U-Net (right) used in teacher-student learning.

impact on the speech enhancement performance.

The rest of paper is organized as follows. In Section II, we provide a brief review of the offline and online Wave-U-Net. In Section III, we introduce the technical details of the proposed online Wave-U-Net with low-overlap window. In Section IV, the effectiveness of the proposed method is demonstrated by speech enhancement experiments simulated as a in-vehicle communication by using measured driving noises. We conclude this paper in Section V.

II. Speech Enhancement with Wave-U-Net

A. Offline Wave-U-Net

In this section, we describe the problem formulation and network structure of single-channel speech enhancement using offline Wave-U-Net [17]. 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, target speech signal, and noise signal, respectively, the aim of offline Wave-U-Net is to estimate the clean speech signal

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

with the whole observed mixture signal $m(1), \ldots, m(T)$. 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 errors (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 upsampling (US) blocks followed by a 1D convolutional layer each, namely, the bottleneck layer and output layer. An illustration of the offline Wave-U-Net architecture is shown as the teacher model in the left of Fig. 1. The DS blocks stack a 1D CNN layer followed by a nonlinear function and then halves the feature resolution by discarding every other feature, and the US blocks stack a US layer and a 1D CNN layer with nonlinearity, where the US layer applies convolution after recovering the intermediate values by interpolation to double the feature resolution. Note that instead of using transposed convolution layers that apply convolution after padding zeros between each original value, performing upsampling with interpolation can avoid aliasing artifacts caused by zeros, which may degrade the enhancement performance. Except for the output layer, where the 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 [21].

B. Online Wave-U-Net

We previously extended the above offline Wave-U-Net to online processing with sequential block-wise speech enhancement [18]. Instead of performing speech enhancement on all mixture signals $\{m(t)\}_{t=1,\ldots,T}$, we perform sequential speech enhancement on a fixed-length input segment $m_i(\tau) = m(iS - S + \tau)\omega_{\text{a}}(\tau)$. Here, $\tau = 1, \ldots, K$ and $i = 1, \ldots, I$ denote the time and segment indices, respectively, and $\omega_{\text{a}}(\tau)$ is an analysis window function with length of K and shift of S. The online Wave-U-net model is trained to estimate clean speech $\tilde{s}_i(1), \ldots, \tilde{s}_i(K) = \mathcal{G}_\phi(m_i(1), \ldots, m_i(K))$ when the mixture segment $m_i(\tau)$ is provided. Here, $\mathcal{G}_\phi(\cdot)$ is a nonlinear function represented by a network with parameter ϕ. Similarly, the estimated noise $\tilde{n}_i(\tau)$ is then obtained by subtracting $\tilde{s}_i(\tau)$ from the mixture.

The most common way to achieve low-latency applications is to shorten the length of the input segment to reduce the algorithmic delay and processing time, where the algorithmic delay is defined as the waiting time to the first processing that is determined by the window length K. However, the available information for inference may also be reduced with a shorter input, leading to a performance decrease. Therefore,

the tradeoff between speech enhancement performance and segment length must be taken into considered. To avoid the speech enhancement performance decrease with short segments, a teacher-student learning method [19] is applied to train the online Wave-U-Net model. Teacher-student learning is a network training technique that transfers the knowledge of a pre-trained teacher model to a student model, which has been applied to various applications. It can be used, for example, to transfer knowledge between different domains for domain adaptation [22], [23], or to transfer knowledge from a deep large network to a shallow small network for model compression [24], or to transfer knowledge from a bidirectional network to a unidirectional network for online interference [25].

We consider to use the pre-trained offline model to apply the teacher-student learning to train the online model, where $\mathcal{F}_\theta(\cdot)$ is the teacher model and the parameter θ is optimized in pre-training and fixed when training the online model. $\mathcal{G}_\phi(\cdot)$ is considered as the student model, where the parameter ϕ is trained using the criterion that minimizes the following loss function:

$$\mathcal{L}_{\text{stu}} = \mathcal{L}^{'} + \beta\mathcal{L}_{\text{diff}}. \tag{4}$$

Here, $\mathcal{L}^{'}$ is the MSE between the signals $\tilde{s}_i(\tau)$ and $\tilde{n}_i(\tau)$ estimated by the student model, and the windowed ground truth signals $s_i(\tau) = s(iS - S + \tau)\omega_{\text{a}}(\tau)$ and $n_i(\tau) = n(iS - S + \tau)\omega_{\text{a}}(\tau)$, which is expressed as

$$\mathcal{L}^{'} = \mathbb{E}\Big[|\tilde{s}_i(\tau) - s_i(\tau)|^2 + |\tilde{n}_i(\tau) - n_i(\tau)|^2\Big]. \tag{5}$$

$\mathcal{L}_{\text{diff}}$ is the MSE between the signals estimated by the student model and those of the teacher model that are segmented by the same analysis window $\omega_{\text{a}}(\tau)$. The second term is expressed as

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\Big[|\tilde{s}_i(\tau) - \hat{s}_i(\tau)|^2 + |\tilde{n}_i(\tau) - \hat{n}_i(\tau)|^2\Big], \tag{6}$$

where $\hat{s}_i(\tau) = \hat{s}(iS - S + \tau)\omega_{\text{a}}(\tau)$ and $\hat{n}_i(\tau) = \hat{n}(iS - S + \tau)\omega_{\text{a}}(\tau)$. $\beta \geq 0$ is a parameter that weighs the importance of the two terms in (4). Fig. 1 shows the teacher-student learning of the online Wave-U-Net.

The student model has a similar network architecture with the teacher model, where the only difference is the number of DS and US blocks, which is defined as L'.

III. Online-Wave-U-Net with Low-Overlap Window

A. Window Processing in Online Wave-U-Net

To perform signal processing smoothly, the input mixture is segmented by an analysis window ω_{a} with a shift length of S. The frames that have been enhanced by Wave-U-Net are then multiplied by the synthesis window function ω_s and added together. This operation is referred to as overlap addition (OLA). To guarantee that the waveform signals segmented by the analysis window can be reconstructed to the original signals, the synthesis window should be chosen to satisfy the following equation,

$$\sum_i \omega_{\text{a}}(\tau)\omega_{\text{s}}(\tau) = 1. \tag{7}$$

This is called the perfect reconstruction property.

Besides the above property, in frequency-domain speech signal processing, window functions such as hanning window and hamming window are generally chosen due to their frequency responses, which are suitable for frequency analysis. The online Wave-U-Net also uses the Hanning window as the analysis window, as in the frequency-domain signal processing. However, as we mentioned above, the window length of Hanning window lower bounds the algorithmic latency, which becomes a limitation to the low-latency applications.

B. Windowing with Low-Overlap Windows

The conventional analysis window for online Wave-U-Net is the Hanning window, which is widely used as an analysis window for frequency-domain speech signal processing. However, using a Hanning window limits the promising ways to reduce the algorithmic delay, which is equivalent to the window length. Since Wave-U-Net applies separation in the time domain, it is not necessary to use the Hanning window, which is commonly chosen because of its appropriate frequency response. In this section, we propose using low-overlap window as an alternative of the Hanning window in online Wave-U-Net to reduce the algorithmic latency while keeping the segment length.

Low-overlap window is a window function applied in the real-time speech compression [20] and used for the modified discrete cosine transform (MDCT). Fig. 2 shows an illustration of the low-overlap window. The window function consists of three parts, namely, the zero region, the overlap region, and the constant region. All values in the zero region are 0, and all values in the constant region are 1. The overlap region uses the value of the following window function used in the Vorbis codec [27]:

$$w(\tau) = \sin\left[\frac{\pi}{2}\sin^2\left(\frac{\pi\left(\tau + \frac{1}{2}\right)}{2D}\right)\right], \tag{8}$$

where D is the length of the overlap region. When the shift length is half of the window length, the low-overlap window satisfies the perfect reconstruction property. When the signal is multiplied by this window function, the information in the zero region disappears, so the zero region information does not need to be prepared in advance and, as a result, the lower bound of the algorithmic delay is the sample length obtained by subtracting the zero region length from the window length. For example, if 25% of the window length is set as the zero-region, the algorithmic delay is reduced by 25% while the frame lengths are the same. The shape of the low-overlap window varies with the length of the zero region as shown in Fig. 3. Since Wave-U-Net applies 1D CNN with zero padding at both end and then downsamples, the samples closer to the center of the frame are considered to have less artifacts due to

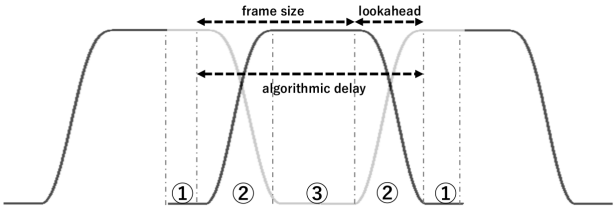

Fig. 2. Low-overlap window.

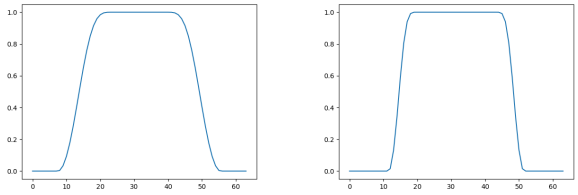

Fig. 3. Change in the shape of the low-overlap window when the length of the zero region is varied: case where the zero region is 25% of the window length (left) and case where the zero region is 40% of the window length (right).

zero padding. This motivates us to expect a small degradation in speech enhancement performance when applying a low-overlap window, which also has zero padding at both ends. Furthermore, the overlap region in the low-overlap window is small, which means the signal region affected by the next frame is small. Therefore, shortening the overlap region can loosen the upper bound of the processing time for the next frame, which is advantageous for the real-time processing.

IV. Experiments

In this section, assuming an in-vehicle communication application, we evaluate the speech enhancement performance and algorithmic delay of online Wave-U-Net using a Hanning window (baseline) or the low-overlap window (proposed) as the analysis window, by simulating an in-vehicle environment with measured driving noise. The SDR, source-to-interferences ratio (SIR), and sources-to-artifacts ratio (SAR) [28] are used as evaluation metrics for speech enhancement performance. We also use the perceptual evaluation of speech quality (PESQ) [29] and short-time objective intelligibility (STOI) [30] as the measures to evaluate the quality of the enhanced speech signal.

A. Datasets

We excerpted utterances of clean speech spoken by 10 speakers from the CMU Arctic database [31], including 100 utterances for each speaker. six speakers (four males and two females) labeled as {“aew”, “ahw”, “aup”, “axb”, “eey”, “fem”} were used to generate the training and validation datasets, and the other speakers (two males and two females) labeled as {“awb”,“bdl”, “clb”, and “slt”} were used for the test. All utterances were about 3 to 7 seconds long. We excerpted noise signals from the JEIDA-NOISE database [32], which were recorded in two different types of car that moved at high speed. There was about 1-hour of data available for each type. We used the 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 by a randomly selected scaling parameter in [0.2, 0.9], while the mixture signals for the test were generated by setting the signal-to-noise ratio (SNR) to {-3, 0, 3} dB. For each SNR, 50 test samples were generated. The average SDR of the input mixture signals was 0.07 dB. All the signals were sampled at 16 kHz.

B. Experimental Setup

For the online Wave-U-Net model, we used a model consisting of 8 DS and US blocks, namely, $L' = 8$. The filter size in the 1D CNN layer was set to 15, 5, 15, and 1 for the DS, US blocks, bottleneck, and output layers, respectively. The number of channels in the lth DS and US blocks was $20l$, and the numbers of channels in the bottleneck and output layers were $20(L+1)$ and 1, respectively. The Adam optimizer with a learning rate of 0.0001 was used for training, and the batch size was set to 32. The trained offline model was used as a teacher model to train a student model with an input segment length of $K = 1,024$. The input segment length of the offline model used as the teacher model was set to $64,000$. The weight parameter β for teacher-student learning was set to 1. For the analysis window, we used three types of low-overlap window, in which the ratio of the zero-region length to the total window length was set to $\{10, 25, 40\}$, and the Hanning window as the baseline. For the synthesis window, we used the optimal synthesis window corresponding to each analysis window. The window length was set to $1,024$ samples, which was the input segment length of the online Wave-U-Net. The shift length of the window function were 512 samples, which was half the window length.

C. Experimental Results

Table I shows the algorithmic delay and average SDR, SIR, SAR, PESQ, and STOI of online Wave-U-Net when using each window function. Among all the window functions, the Hanning window used in the baseline system achieved the highest speech quality in terms of SDR. The proposed method achieved a high SDR of more than 14 dB in all cases. However, we found that scores tended to decrease as the ratio of the zero region used in the low-overlap window increased, indicating a tradeoff between both the speech quality and separation performance and the ratio of the zero region. Comparing the results using the Hanning window and the low-overlap window with a zero region length of 40% of the window length, which had the lowest algorithmic delay in the experiments, we found that the proposed method reduced the algorithmic delay of online Wave-U-Net by 40% while keeping the SDR reduction to 1.15 dB. Reverberation is generally considered to affect the source separation performance, and is one of the important

TABLE I
COMPARISON OF AVERAGE SDR [DB], SIR [DB], SAR [DB], PESQ, AND STOI WITH DIFFERENT WINDOW FUNCTIONS.

Analysis window ω_a	Ratio of zero area [%]	Algorithmic delay [ms]	SDR	SIR	SAR	PESQ	STOI
Low-overlap window	10	57.6	**15.15**	**27.64**	**15.56**	**3.18**	**0.95**
	25	48.0	**14.73**	**26.91**	**15.17**	**3.14**	**0.95**
	40	38.4	**14.19**	**26.33**	**14.62**	**3.10**	**0.94**
Hanning window		64.0	**15.34**	**27.08**	**15.80**	**3.19**	**0.95**

factors to be considered in the future. Since the reverberation in a car is relatively small, the proposed system is expected to work even in the presence of reverberation.

V. CONCLUSION

In this paper, we reconsidered the window processing in the conventional online Wave-U-Net, and proposed the use of a low-overlap window as an alternative to the Hanning window to reduce the algorithmic delay. Through evaluation experiments simulating the interior of a car being driven, it was found that the proposed method reduced the algorithmic delay of online Wave-U-Net by 40% while keeping the SDR reduction to 1.15 dB.

ACKNOWLEDGMENT

This work was partly supported by JSPS KAKENHI Grant Number 19H04131, and JST CREST JPMJCR19A3.

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*, 26(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*, 23(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] S. Nakaoka, L. Li, S. Inoue, and S. Makino, "Teacher-student learning for low-latency online speech enhancement using Wave-U-Net," in *Proc. ICASSP*, pp. 661-665, 2021.
[19] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NIPS*, 2014.
[20] Valin, J. M., Terriberry, T. B., Montgomery, C., and Maxwell, G. "A High-Quality Speech and Audio Codec With Less Than 10 ms Delay," *IEEE Trans. ASLP*, 18(1), pp. 58–67, 2009.
[21] 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.
[22] Z. Meng, J. Li, Y. Zhao, and Y. Gong, "Conditional teacher-student learning," in *Proc. ICASSP*, pp. 6445–6449, 2019.
[23] 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.
[24] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," in *Proc. ICLR*, 2018.
[25] 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.
[26] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform, " in *IEEE Trans. ASSP,* 32(2) pp. 236–243, 1984.
[27] Vorbis I specification, accessed July 14, 2021, http://www.xiph.org/vorbis/doc/Vorbis_I_spec.html, 2004.
[28] 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.
[29] 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.
[30] 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.
[31] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. SSW*, pp. 223–224, 2004
[32] JEIDA Noise Database, accessed Oct 21, 2020, http://research.nii.ac.jp/src/en/JEIDA-NOISE.html