Stacked U-Net with High-level Feature Transfer for Parameter Efficient Speech Enhancement

Jinyoung Lee and Hong-Goo Kang
Yonsei University, Seoul, South Korea
E-mail: jylee@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr

***Abstract*—In this paper, we present a stacked U-Net structure-based speech enhancement algorithm with parameter reduction and real-time processing. To significantly reduce the number of network parameters, we propose a stacked structure in which several shallow U-Nets with fewer convolutional layer channels are cascaded. However, simply stacking the small-scale U-Nets cannot sufficiently compensate for the performance loss caused by the lack of parameters. To overcome this problem, we propose a high-level feature transfer method that passes all the multi-channel output features, which are obtained before passing through the intermediate output layer, to the next stage. Furthermore, our proposed model can process analysis frames with short lengths because its downsampling and upsampling blocks are much smaller than the conventional Wave U-Net method; theses smaller layers make our proposed model suitable for low-delay online processing. Experiments show that our proposed method outperforms the conventional Wave U-Net method on almost all objective measures and requires only 7.21% of the network parameters when compared to the conventional method. In addition, our model can be successfully implemented in real time on both GPU and CPU environments.**

I. Introduction

Speech enhancement, which attempts to obtain a clean target speech signal in a noisy environment, plays a crucial role in speech communications and voice-controlled interfaces such as automatic speech recognition (ASR) [1], [2]. Recently, deep learning-based speech enhancement methods have demonstrated significantly improved performance compared to conventional unsupervised statistics-based methods. However, since deep learning-based methods require high computational complexity and large amounts of memory, they are difficult to implement on embedded devices, which normally have limited capacity. In order to further expand application areas, it is necessary to reduce the number of computations and memory usage so that these methods can be processed using low-capacity devices. However, it is very challenging to implement a neural network-based speech enhancement system that achieves high performance on such a low-capacity device.

At an early stage of deep learning-based speech enhancement, most methods attempted to estimate time-frequency masking values using various types of network architectures under supervised training frameworks [3], [4], [5], [6]. These methods first estimated clean magnitudes by multiplying the estimated masking values to noisy magnitudes, then reconstructed enhanced time-domain signals using noisy phases and inverse short-time Fourier transform (iSTFT). In other words, they substituted only conventional statistics-based gain estimation modules with neural network architectures, but their frequency-domain processing frameworks, which use noisy phase terms, remained unchanged. Signals reconstructed using noisy phase terms are prone to distortion and generally have low intelligibility.

To solve the problems caused by noisy phase terms, end-to-end training methods that directly process the input of raw time-domain waveforms have been proposed. SEGAN [7] and Wave U-Net [8] are the most popular methods that have implemented this strategy. Both methods use a U-Net structure that can be trained in the end-to-end manner using an encoder and a decoder module. The encoder generates multi-scale features by gradually decreasing the input's time scales, i.e., downsampling, while increasing the number of channels in the convolutional layers; the decoder generates high-resolution signals by merging the generated features with skip connections passed from the encoder module. Although many researchers have studied and improved the U-Net structure [9], [10], [11], [12], [13], these types of encoders exponentially increase the number of channels as the the layer depth increases, which significantly increases the number of network parameters. Also, since the downsampling processes halve the length of the input signals, it is necessary to use input signals with long lengths to use the deep encoding layers. In other words, this strategy is not suitable for implementing real-time systems with low latency.

In this paper, we aim to drastically reduce the number of parameters in the U-Net structure while maintaining the quality of the enhanced speech signals. To reduce the number of parameters, instead of using one large U-Net, we propose a stacked U-Net that consists of several cascaded small-scale U-Nets. Since each small-scale U-Net uses shallow encoding and decoding layers, we can significantly reduce the number of channels and network parameters. In addition, the stacked U-Net structure generates high-resolution signals by iteratively combining multi-scale features, which progressively refines noisy signals through multiple encoding and decoding processes. While the stacked structure has been already investigated for time-frequency domain speech enhancement [14], [15], most implementations simply stacked the existing U-Net structure without analysis of how effective the stacked structure is when using the same number of layers as the single U-Net structure, and how to connect each of the small-scale U-Net to be the most effective.

In this paper, we investigate the efficient structure of the

stacked U-Net by analyzing different connection methods and different weights for intermediate supervisions. In addition, we propose a high-level feature transfer (HFT) method that transmits the well-analyzed outputs of former stages to the next stage's inputs, which gradually improves the performance as the stages are stacked. The proposed HFT method transfers multi-channel features generated before the previous stage's output layer to the next stage, rather than using a single-channel signal in which the information is compressed due to the intermediate output layer. Compared to various types of connection methods, the proposed HFT method proves to be the most effective for stacking.

The proposed stacked U-Net model also has the following advantages. First, the conventional stacked U-Net approaches [15], [16] generally forcibly limit the performance of each stage for progressive learning, in such a way as using weighted loss. The proposed method enables progressive learning through the HFT method as well as equal weighting at each stage and does not forcibly limit the performance at former stages. Therefore, if the user device has insufficient capacity, we can further reduce the number of parameters and complexity of the proposed model by disconnecting the latter stages and using only the remaining former stages so that the performance penalty is much less than that for conventional approaches. In addition, the proposed model enables low-delay online processing using short input signals.

II. Related Work

Wave U-Net is an end-to-end source separation model that directly utilizes time-domain signal as an input [8]. As shown in the upper part of Fig. 1, Wave U-Net consists of an encoder, a decoder, and skip connections that pass each layer's encoder output to the decoder module. The encoder and the decoder include one-dimensional convolutional layers with downsampling and upsampling blocks, respectively. The downsampling blocks generate high-level features while gradually increasing the number of channels, and the upsampling blocks generate multi-resolution features and then combine these features with features transferred over skip connections to predict the desired high-resolution signals.

Stacked U-Net repeats the encoder and decoder processing several times using a cascaded structure to gradually refine input signals. Shah et al. [17] used this cascaded structure for image segmentation to generate high-resolution pixel-level images; this strategy resulted in higher performance than the conventional U-Net structure with fewer parameters.

Deep SEGAN (DSEGAN) [16] is a multi-stage speech enhancement network based on SEGAN [7] that uses generative adversarial networks for end-to-end time-domain speech enhancement. Each generator improves its predecessor's output to progressively refine noisy signals. However, this method has inherent limitations because the network requires more parameters and a great number of computations as the number of stages increases due to the simple stacking of SEGAN generators. PL-CRNN is a progressive learning framework that uses convolutional recurrent neural networks (CRNNs) [15]. In order to reduce the number of parameters, this model used multiple CRNN sub-nets [18] where the number of channels is reduced and the LSTM layers share parameters with each other. The progressive learning method and dense connections that combine the outputs from different stages were used to overcome the performance gap caused by the downsized network.

However, both DSEGAN and PL-CRNN regulated the influence of each stage forcibly during the progressive learning stage. They gave different weights to each stage's loss value, and set the intermediate signal-to-noise ratio (SNR) improvement level heuristically. Therefore, these models limit the degree of performance improvement in each stage, which makes it hard to flexibly control the number of stages when used in the inference. To enhance the system's flexibility in response to device capacity, it is important to implement a method for reliably adjusting the number of enhancement stages.

III. Proposed system description

In this section, we describe the structure of the proposed stacked U-Net, which comprises several small-scale U-Nets. This algorithm drastically reduces the number of network parameters when compared to the baseline deep U-Net structure such as that used in Wave U-Net. To further improve the performance, we also propose a high-level feature transfer method that transfers the multi-channel feature map output from the previous stage to the next stage to minimize information loss.

A. Stacked U-Net structure

Fig. 1 shows the structure of Wave U-Net and our stacked U-Net. Both models are similar in that the encoder halves the feature map's time step in each downsampling block, and the decoder upsamples the feature map resolution using linear interpolation in each upsampling block. However, the two models are different because Wave U-Net has a single deep U-Net structure but stacked U-Net has a structure in which several shallow U-Nets are cascaded. In addition, Wave U-Net uses 12 blocks in both the encoder and decoder but one stage in stacked U-Net consists of smaller blocks. In this paper, we set the total number of blocks to be equal to that of Wave U-Net in order to provie a fair performance comparison. For example, when using three U-Nets for stacked U-Net as shown in Fig. 1, each encoder and decoder has four blocks for a total of 24 blocks.

Table I provides a detailed comparison of the network structures. Although the total number of blocks is the same, the number of parameters required by stacked U-Net is much smaller than that required by Wave U-Net because the number of feature map channels is smaller due to the shallow U-Net structure. In addition, we can greatly reduce the analysis frame length because we only need a small number of downsampling processes, which enables the implementation of a real-time system with low delay. Note that Wave U-Net is not suitable for low-delay real-time processing because of the large number of downsampling blocks.

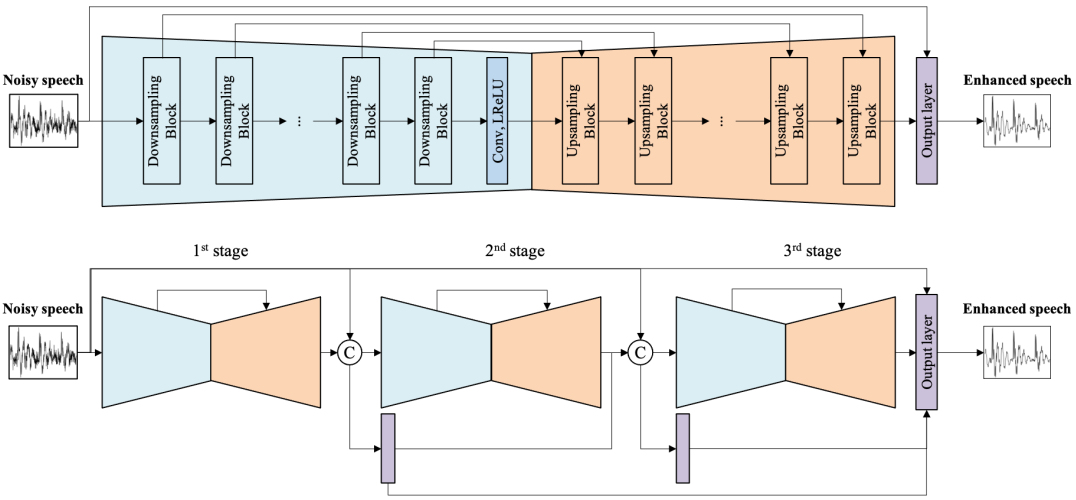

Fig. 1. Illustration of Wave U-Net (top) and the proposed stacked U-Net model (bottom), which has three stages. © denotes a channel-wise concatenation.

To maximize the advantages of the stacked U-Net structure, it is important to effectively transfer key information analyzed in the previous stage to the next stage. We implement various types of connection methods between two consecutive small-scale U-Nets, and propose an optimal method that improves the stacking efficiency.

B. High-level feature transfer

Most of the conventional stacked U-Net systems simply pass the single-channel output from the previous stage to the next stage's input. However, the single-channel output, which has been compressed from the multi-channel features in the previous output layer, is reanalyzed back to the multi-channel features in the next stage. This reanalysis is a redundant process, and the previously well-analyzed information cannot be fully utilized. In other words, we need a method that fully utilizes the information analyzed in the previous stage.

In the baseline U-Net structure, the final multi-channel features, which are generated before the output layer, have the same length as the output signal. Therefore, it is beneficial to pass these uncompressed high-level multi-channel features directly to the next stage's input because the features preserve all the relevant information; this preservation reduces the next stage's burden. Even this setup, we are still able to obtain the intermediate enhanced output at each stage by separately applying the intermediate output layer. Additionally, the dense connection is applied to the output layers to concatenate all of the previous intermediate outputs and help generate the next stage output.

IV. Experiments

A. Data setup

All experiments are performed using the database [19], which was made of Voice Bank corpus [20] and Diverse Environments Multichannel Acoustic Noise Database (DEMAND) [21]. The database has been widely used in previous studies [7], [16]. The noisy database used for training was created by mixing clean signals uttered by 28 speakers from the Voice Bank corpus with ten different noise types (two artificially generated noises and eight noises from DEMAND) and four SNR values (15, 10, 5, and 0 dB). The entire training set consists of 11,572 utterances, and among them, 1,000 randomly selected utterances were used for validation, while the remaining 10,572 utterances were used for training.

TABLE I
Structural comparison of Wave U-Net and the proposed stacked U-Net

	Wave U-Net	Stacked U-Net
Number of U-Nets	1	3
Number of blocks in one encoder	12	4
Total number of blocks	24	24
Channel increment per downsampling block	24	16
Number of parameters	10.26 M	0.74 M

The noisy database used for testing was created by mixing clean signals uttered by two additional speakers from the Voice Bank corpus with five unseen noises from DEMAND and four SNR values (17.5, 12.5, 7.5, and 2.5 dB). The test set consists of 824 utterances. All data were originally recorded at a sampling frequency of 48 kHz, and we resampled them to 16 kHz for processing.

B. Network setup

Our proposed network comprises three cascading U-Nets, and each U-Net has four downsampling and upsampling blocks. The kernel size of the 1-D convolutional layer is 15 for the downsampling blocks and bottleneck layer, 5 for the upsampling blocks, and 1 for the output layer. The number of channels starts at 16 and increases by 16 as the encoder layer becomes deeper, and decreases by 16 in the decoder. Every convolutional layer is followed by LeakyReLU activation

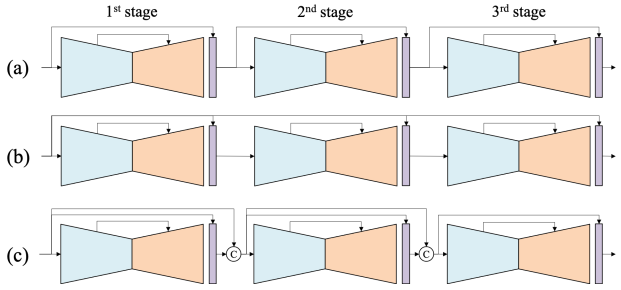

Fig. 2. Conventional single-channel connection methods. (a) Model A: baseline connection. (b) Model B: noisy connection. (c) Model C: dense connection.

except the output layer, which is followed by hyperbolic tangent activation. The structural differences from Wave U-Net are specified in Table I. The network was trained with a batch size of 16 using the Adam optimizer [22] with a learning rate of 0.0001, and was initialized using the Glorot normal initializer [23]. The input length to the network was set to 16,384 samples to provide a fair comparison with the baseline and 512 samples for low-delay real-time processing. We added intermediate supervisions with MSE losses between the intermediate outputs and the clean target without needing to provide each stage with different weights: $L = \sum_{n=1}^{N} \frac{1}{N} MSE(y, \hat{y}_n)$, where y is the clean target speech, $\hat{y}_n$ is the intermediate output of the n^{th} stage, and N is the number of stages.

C. Model comparison

To determine the impact of the high-level feature transfer method, we compared the proposed method with several connection methods previously used for single-channel signals. Fig. 2(a) is a baseline method that simply concatenates multiple U-Nets. The subsequent stage receives the previous single-channel output as an input and sends it to the output layer via skip connections. Fig. 2(b) transfers the noisy signal to all of the stages' output layers using skip connections to provide raw information. Fig. 2(c) uses a dense connection method that concatenates the noisy signal and entire previous outputs; these concatenated features are then propagated to the next stages' input and output layer.

Table II and Fig. 3 present the compared performance of Wave U-Net and various stacked U-Net models with different connection methods. Our comparisons use five objective measurements: composite measures for signal distortion (CSIG), noise distortion (CBAK) and overall speech quality (COVL), the perceptual evaluation of speech quality (PESQ), and segmental SNR (SSNR)[1]. In Fig. 3, the x-axis represents the stages' index and the y-axis represents the score values. After end-to-end training on the stacked U-Net consisting of three stages, we measured the performance of the enhanced output at each stage to ensure that the performance improved gradually.

[1]https://www.crcpress.com/downloads/K14513/K14513_CD_Files.zip

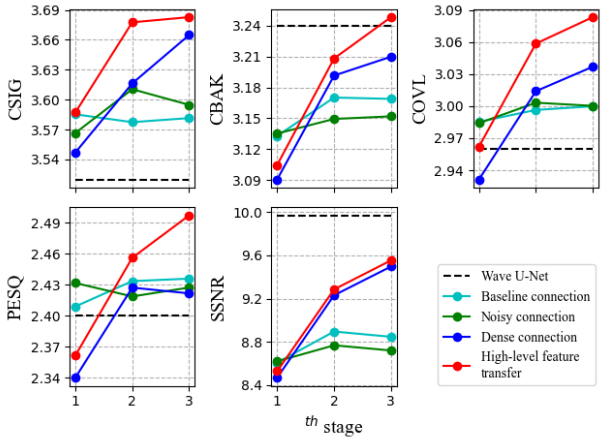

Fig. 3. Performance comparison of several connection methods.

The baseline connection model (A) and noisy connection model (B) exhibit low performance that does not gradually increase when passing through the stages. The dense connection model (C) aggregates the information at different stages and its performance improves as it passes through the stages, but its PESQ score is poor. The proposed HFT method outperforms other connection methods on all objective measures and its performance improves gradually. Transferring multi-channel features rather than transferring a single-channel signal is beneficial as it gives more various information to subsequent networks and maximizes the stacking efficiency. In addition, the proposed model is better than the conventional Wave U-Net on all objective measures except SSNR while using only 7.21% of the parameters of Wave U-Net. Also, the inference up to the second stage of the proposed model only requires 66.47% of the number of parameters and computations compared to the inference up to the third stage, but the output performance is only 1.33% lower. Therefore, when operating on a device whose capacity is too low to use three stages, the complexity can be greatly reduced using only two stages; however, performance degradation is still relatively small.

D. Effect of the number of stages

In this section, we analyze the performance variation by gradually increasing the number of stacked U-Net stages. Fig. 4 presents the performance of the intermediate outputs obtained by stacking N stages. The proposed model demonstrates the gradual performance improvement on all performance-related scores until the fourth stage is stacked. However, when stacking five stages, the performance of the late outputs deteriorated. It seems that after stacking approximately 50 convolutional layers, the network becomes too deep, which impedes the gradient flow.

The number of included stages will differ depending on the user device's computing level. When adapting to user devices, there is no need to train all of the models separately, changing the number of stages. The proposed stacked U-Net model can

TABLE II
OBJECTIVE MEASURES OF THE PROPOSED MODELS AGAINST WAVE U-NET

Model	CSIG	CBAK	COVL	PESQ	SSNR
Wave U-Net [24]	3.52	**3.24**	2.96	2.40	**9.97**
Stacked U-Net (Model A)	3.58	3.17	3.00	2.44	8.85
Stacked U-Net (Model B)	3.60	3.15	3.00	2.43	8.72
Stacked U-Net (Model C)	3.65	3.21	3.04	2.42	9.50
Stacked U-Net HFT (Proposed)	**3.69**	**3.24**	**3.08**	**2.49**	9.52
Stacked U-Net HFT (RT)	3.69	3.20	3.04	2.40	9.43

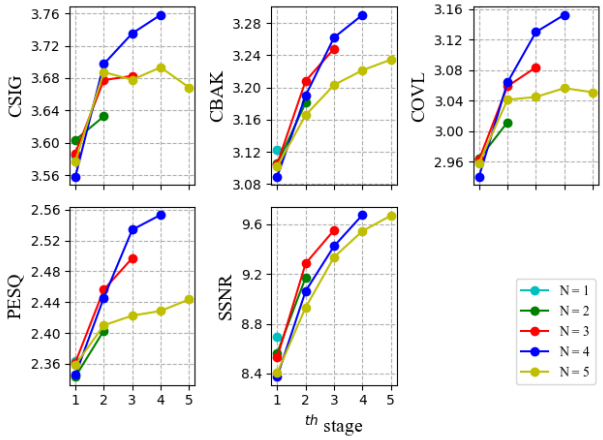

Fig. 4. Performance comparison according to the number of stages.

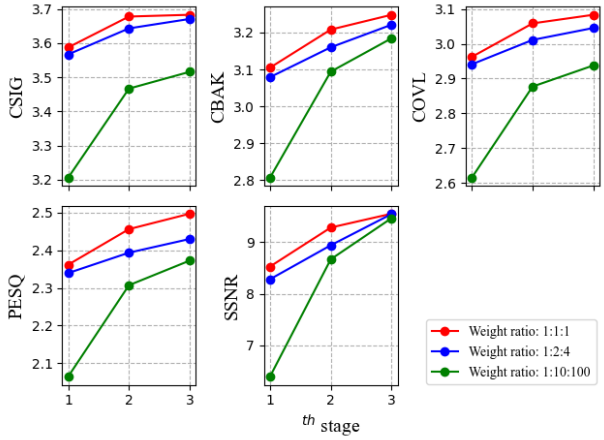

Fig. 5. Performance comparison according to the weight ratios.

be trained up to four stages at once, and can be cut to the required number of stages.

E. Effect of the weighted loss ratio

In this section, we analyze the performance variation in the case that we assign different weighting values when calculating the intermediate supervision losses. The conventional stacked U-Net approaches introduced weighted loss criteria in each stage: $L = \sum_{n=1}^{N} \alpha_n MSE(y, \hat{y}_n)$ [15], [16]. We performed similar experiments that examine whether assigning different weighting values truly benefits the final performance in a progressive learning structure. The weight ratios of α_n were set to 1:2:4, 1:10:100 and the proposed 1:1:1, when $\sum_{n=1}^{N} \alpha_n = 1$. As shown in Fig. 5, assigning a small weight to former stages only reduced the performance of former outputs, but also did not help to improve the final performance. As a result, we assigned the same weights to all stages.

F. Real-time processing

We reduced the analysis frame length of the input signals to 32 ms (512 samples) to implement a real-time system with low latency. To minimize discontinuity at the frame boundary, we also apply a 50% overlap-and-add method using a Hanning window in the inference. We measured a real-time factor (RTF), which is defined as the ratio between the processing time and the length of an input utterance [25]. The RTF was 0.19 when tested using Nvidia GeForce RTX 2080 Ti GPU and 0.54 when tested using dual-core i5-7360U CPU, which confirms that our proposed model can be used in real-time environments. Also, as shown in Table II, the performance of the proposed real-time processing model (Stacked U-Net HFT (RT)) is not significantly affected despite its inability to use long contextual information.

V. CONCLUSION

In this paper, we proposed a stacked Wave U-Net model with a progressive learning framework for speech enhancement. Each stage's U-Net architecture is designed to have a small number of layers and channels such that the overall number of network parameters can be drastically reduced. To overcome the performance loss caused by parameter reduction, we proposed a high-level feature transfer method that passes multi-channel information from the previous stage's output to the next stage's input. In addition, the system predicts the intermediate enhanced outputs in the separately applied output layers, and these predicted outputs are passed to the output layers of the latter stages; this information provision helps generate the next stage's output by aggregating information from the previous stages. Experimental results showed that our proposed stacked Wave U-Net model successfully reduces the number of network parameters and improves the enhancement performance. Since the input signals' frame length is fairly small, our model is also suitable for real-time processing.

ACKNOWLEDGMENT

This project was supported by the Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (No. 2019-0-01558: Study on audio, video, 3d map and activation map generation system using deep generative model)

REFERENCES

[1] C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A Scalable Noisy Speech Dataset and Online Subjective Test Framework," in *Proc. Interspeech 2019*, 2019, pp. 1816–1820. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-3087

[2] C. Zorilă, C. Boeddeker, R. Doddipatla, and R. Haeb-Umbach, "An investigation into the effectiveness of enhancement in asr training and test for chime-5 dinner party transcription," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 47–53.

[3] Y. Li and D. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Communication*, vol. 51, no. 3, pp. 230–239, 2009.

[4] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7092–7096.

[5] 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, 2018.

[6] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 708–712.

[7] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," in *Proc. Interspeech 2017*, 2017, pp. 3642–3646.

[8] S. E. Daniel Stoller and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 2018, pp. 334–340.

[9] R. Giri, U. Isik, and A. Krishnaswamy, "Attention wave-u-net for speech enhancement," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 249–253.

[10] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," *arXiv preprint arXiv:1903.03107*, 2019.

[11] E. T. Kaspersen, T. Kounalakis, and C. Erkut, "Hydranet: A real-time waveform separation network," in *ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4327–4331.

[12] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech 2020*, 2020, pp. 2472–2476.

[13] A. Défossez, G. Synnaeve, and Y. Adi, "Real Time Speech Enhancement in the Waveform Domain," in *Proc. Interspeech 2020*, 2020, pp. 3291–3295.

[14] S. Abdulatif, K. Armanious, K. Guirguis, J. T. Sajeev, and B. Yang, "Aegan: Time-frequency speech denoising via generative adversarial networks," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 451–455.

[15] A. Li, M. Yuan, C. Zheng, and X. Li, "Speech enhancement using progressive learning-based convolutional recurrent neural network," *Applied Acoustics*, vol. 166, p. 107347, 2020.

[16] H. Phan, I. V. McLoughlin, L. Pham, O. Y. Chén, P. Koch, M. De Vos, and A. Mertins, "Improving gans for speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 1700–1704, 2020.

[17] S. Shah, P. Ghosh, L. S. Davis, and T. Goldstein, "Stacked u-nets: A no-frills approach to natural image segmentation," *arXiv preprint arXiv:1804.10343*, 2018.

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

[19] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech," in *9th ISCA Speech Synthesis Workshop*, pp. 146–152.

[20] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013, pp. 1–4.

[21] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, no. 1. Acoustical Society of America, 2013, p. 035081.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

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

[25] A. V. Ivanov, P. L. Lange, D. Suendermann-Oeft, V. Ramanarayanan, Y. Qian, Z. Yu, and J. Tao, "Speed vs. accuracy: Designing an optimal asr system for spontaneous non-native speech in a real-time application," *Proc. of the IWSDS, Saariselk, Finland*, 2016.