INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

NAAGN: Noise-aware Attention-gated Network for Speech Enhancement

Feng Deng, Tao Jiang, Xiao-Rui Wang, Chen Zhang, Yan Li

Kuai Shou Technology Co., Beijing, China

{dengfeng, jiangtao, wangxiaorui, zhangchen03, liyan}@kuaishou.com

Abstract

For single channel speech enhancement, contextual information is very important for accurate speech estimation. In this paper, to capture long-term temporal contexts, we treat speech enhancement as a sequence-to-sequence mapping problem, and propose a noise-aware attention-gated network (NAAGN) for speech enhancement. Firstly, by incorporating deep residual learning and dilated convolutions into U-Net architecture, we present a deep residual U-net (ResUNet), which significantly expand receptive fields to aggregate context information systematically. Secondly, the attention-gated (AG) network is integrated into the ResUNet architecture with minimal computational overhead while furtherly increasing the long-term contexts sensitivity and prediction accuracy. Thirdly, we propose a novel noise-aware multi-task loss function, named weighted mean absolute error (WMAE) loss, in which both speech estimation loss and noise prediction loss are taken into consideration. Finally, the proposed NAAGN model was evaluated on the Voice Bank corpus and DEMAND database, which have been widely applied for speech enhancement by lots of deep learning models. Experimental results indicate that the proposed NAAGN method can achieve a larger segmental SNR improvement, a better speech quality and a higher speech intelligibility than reference methods.

Index Terms: speech enhancement, attention-gated network, residual learning, dilated convolution

1. Introduction

Speech enhancement is one of the most important and challenging tasks in speech applications. And in the last several decades, a large number of speech enhancement approaches have been proposed, including the Wiener filtering method [1], the spectral-subtraction method [2] [3] and statistical-model-based methods [4] - [6], and so on. However, these kinds of speech enhancement methods generally do not perform well in adverse noise environments.

Recently, due to the advances in deep learning, deep neural network (DNN)-based speech enhancement approaches [7] - [11] have been attracting large attention, and the speech enhancement task has been obtained significant improvements. For DNN-based speech enhancement methods, most have mainly focused on estimating the magnitude spectrogram of speech while reusing the phase from noisy speech for reconstruction [7][9]. As we know, to obtain accurate magnitude spectrogram estimation, the temporal contexts are very important. However, conventional DNNs cannot leverage long-term contexts. Therefore, recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have been used for speech enhancement [8] [9] [11]. [8] proposed an RNN model with four hidden LSTM layers, their experimental results show that the RNN model performs better than DNN-based model. [9] developed a CNN model based on dilated convolutions, in which the speech enhancement task is treated as a sequence-to-sequence mapping and the dilated convolutions are used to leverage contexts. Compared with the LSTM model in [8], the model in [9] shows better enhancement performance. [12] presented the U-Net structure for audio source separation tasks, which is a well-known architecture composed as a convolutional encoder-decoder with skip connections. Furthermore, the U-Net has been shown to be also effective to speech enhancement task [13] [14]. Motivated by its good performance, we used the U-Net architecture as the basis for the work presented in this paper.

In this paper, to aggregate long-term contexts, we also formulate the speech enhancement task as a sequence-to-sequence mapping, and propose a deep noise-aware attention-gated network (NAAGN) by introducing residual learning [15], dilated convolutions [9] [10] and attention gate mechanism [16] into the U-Net architecture. The dilated convolutions expand receptive fields compared with conventional convolutions, and the receptive field is a region in the input space that affects a high-level feature. With the formulation of speech enhancement as a sequence-to-sequence mapping, large receptive fields of the NAAGN amount to long-term contexts. The residual blocks are summated to yield high-level features, which preserve and integrate the knowledge learned by all the stacked blocks of ResUNet. The attention gate (AG) block is incorporated into the proposed ResUNet architecture to highlight salient features that are passed through the skip connections, which can increase the long-term contexts sensitivity and prediction accuracy with minimal computational overhead. In addition, a novel noise-aware multi-task loss function, called weighted mean absolute error (WMAE) loss, is proposed, in which both speech estimation loss and noise prediction loss are considered, and noise prediction loss term is added and expected to be complementary to speech estimation. In this way, the NAAGN can balance well between removing amount noise and reducing speech distortion. The experimental results show that the proposed NAAGN method achieves state-of-the-art performance, which outperforms the reference methods over several different metrics.

The remainder of this paper is organized as follows. In Section 2, the proposed NAAGN architecture is described. The performance evaluation is presented in Section 3, and Section 4 gives the conclusions.

2. NAAGN architecture

In this section, we will discuss the details on the proposed NAAGN model. starting with the overview of NAAGN architecture, followed by the sub-blocks of the model. Finally, we will introduce a new noise-aware multi-task loss function to optimize the model, which plays a critical role for accurate speech estimation.

http://dx.doi.org/10.21437/Interspeech.2020-1133

2.1. Overview

The block diagram of the proposed NAAGN method is given in Figure. 1, which includes two parts, the upper part is signal processing, and the lower part is NAAGN architecture. For the signal process procedure, firstly, the input time-domain waveform signal is transformed into time-frequency (T-F) domain by the short-time Fourier transform (STFT) and represented by a magnitude spectrogram and phase information. Secondly, based on the magnitude spectrogram, a multiplicative mask is predicted with NAAGN model. Thirdly, the modified magnitude spectrogram is obtained by applying the multiplicative mask to the input magnitude spectrogram. Finally, combined the modified magnitude spectrogram and input phase spectrogram, the inverse STFT (ISTFT) is applied to obtain the real-valued time-domain waveform. The key problem is how to model and predict the mask accurately, so we emphatically introduce the NAAGN architecture in the following sections.

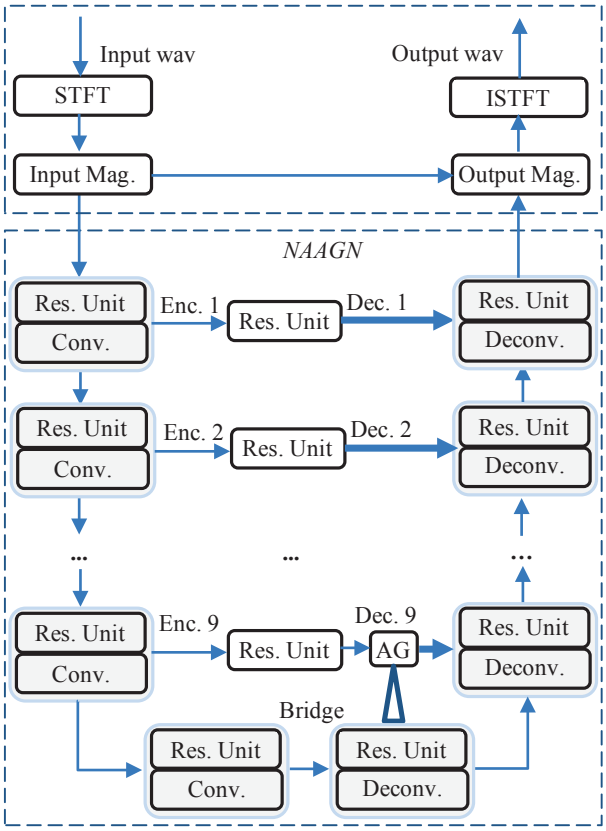

Figure 1: *The block diagram of the proposed NAAGN method. Thick arrows denote concatenated operations, the triangle represents the gating signal for AG block.*

The NAAGN is a variant U-Net architecture applied in STFT domain, and the modifications are summarized as follows. Ahead each convolutional layer of U-Net, a dilated residual (Res.) unit is added, which combines strengths of both dilated convolutions and residual learning. The details of Res. unit will be presented in subsection 2.2. Furthermore, as we know, usually a minor linear shift in the spectrogram has disastrous effects on perception. Thus, the skip connections between layers at the same hierarchical level in the encoder and decoder, are also replaced with Res. units. This furtherly improves low-level information to flow from the high-resolution input to the high-resolution output. Here, we call the variant of U-Net as ResUNet. For the ResUNet, the convolution/deconvolution units (Conv./Deconv.) are also made a few modifications, which can be shown as Figure 2. The convolution kernels are set to be independent to each other by initializing the weight tensors as unitary matrices for better generalization and fast learning [14]. Batch normalization (BN) is implemented on every convolutional/deconvolution layer, and the activation function of ReLU is replaced with leaky ReLU, which yields training more stable. In the encode stage, instead of using max pooling, the convolution operations with different stride sizes over time and frequency directions to prevent spatial information loss. In the decode stage, strided deconvolution operations are used to restore the size of input. In addition, we apply the attention-gated network [16] to the Dec. 9 layer of ResUNet to better exploit context information, and the gate signal is the output of bridge layer, which is denoted by the triangle in Figure 1. Note that in the last layer of the NAAGN model, the batch normalization and leaky ReLU activation were not used and non-linearity function was applied for mask instead. Next, we will introduce the main blocks of modification in Figure 1.

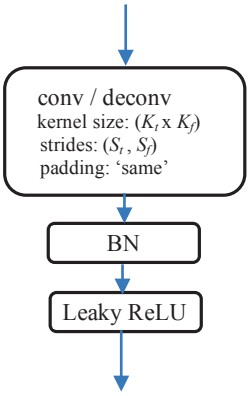

Figure 2: *Schematic of Conv./Deconv. Block. K_f and K_t are the kernel size of convolution filter along the frequency and time axis, respectively. S_f and S_t denote the stride size of convolution filter along the frequency and time axis, respectively. The 'same' results in padding the input such that the output has the same length as the original input. Note that the symbol conv/deconv denotes the convolutional/deconvolution operation, which is different from symbol of Conv./Deconv.*

2.2. Dilated residual unit

In convolutional neural networks, contextual information is augmented typically through the expansion of the receptive fields. One way to achieve this goal is to increase the network depth, which decreases computational efficiency and typically results in vanishing gradients [9]. Another way is to enlarge the kernel size, which likewise raises computational burden and training time. To address this problem, the dilated convolutions were used for multi-scale context aggregation in speech enhancement [9] [10], which are based upon the fact that dilated convolutions can exponentially expand receptive fields without losing resolution or coverage.

To alleviate the vanishing gradient problem, the deep residual learning framework was developed [15], which lead to ease training of the network by improving the propagation of

information and gradients throughout the network. In this paper, motivated by the strengths of both dilated convolutions and residual learning, we present a residual unit with dilated convolutions, and call Res. Unit for short. As shown in Figure 3, the Res. Unit consist of two convolution blocks and an identity mapping. Each convolution block includes a dilated convolution layer, a BN and a leaky ReLU activation function. And the dilation of dilated convolution is applied to both the frequency direction and the time direction, which can aggregate contextual information over both time and frequency dimensions. The identity mapping with Conv. Block of 1x1convolution connects input and output of the unit, which is only used to ensure the same dimensions of two tensors that pass to addition operation.

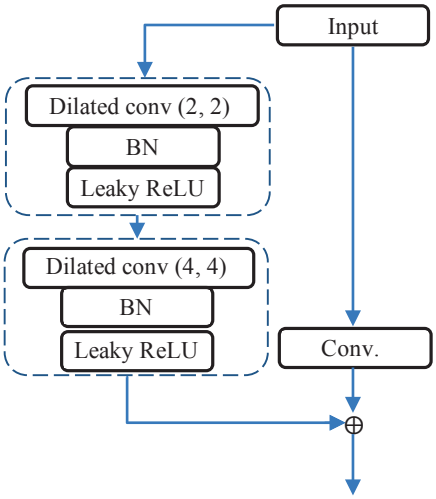

Figure 3: *Schematic of dilated Res. Unit. The tuples (2,2) and (4,4) are dilation rate, which means the dilation applied both in time and frequency directions.*

2.3. Attention gate (AG)

In [16], attention gate (AG) model was introduced for medical imaging that can automatically learn to focus on target structures of varying shapes and sizes and to suppress irrelevant background regions in an input image while highlighting salient features useful for a specific task. Therefore, we incorporate the AG block into our proposed deep ResUNet architecture to better exploit contextual information. Meanwhile, by balancing the computational complexity and performance, we integrated AG into the last skip connection with minimal computational overhead, see Figure 1, to increase the model sensitivity and prediction accuracy. For the AG block, the gating signal is the output of bridge layer, which is used to prune irrelevant and useless lower-level feature responses in skip connection. In this way, only relevant activations can be merged by the concatenation operation.

The schematic of AG unit is shown as Figure 4, in which additive attention [16] is used. By summarizing the AG unit, we can obtain the formulations as follows:

$$\boldsymbol{\varphi} = \boldsymbol{W}_p\sigma_1\left(\boldsymbol{W}_f\boldsymbol{f} + \boldsymbol{W}_g\boldsymbol{g} + \boldsymbol{b}_g\right) + \boldsymbol{b}_p \quad (1)$$

$$\boldsymbol{w} = \boldsymbol{W}_s(\sigma_2(\boldsymbol{\varphi}(\boldsymbol{f}, \boldsymbol{g}; \boldsymbol{\Theta}_{att}))) \quad (2)$$

$$\hat{\boldsymbol{f}} = \boldsymbol{w}\boldsymbol{f} \quad (3)$$

where $\boldsymbol{\Theta}_{att}$ is a set of parameters of AG. $\boldsymbol{f}$ corresponds to the input lower-level feature in skip connection. $\boldsymbol{g}$ is the gating signal contains contextual information to determine focus regions of input feature $\boldsymbol{f}$. $\boldsymbol{W}_g$, $\boldsymbol{W}_f$ and $\boldsymbol{W}_p$ are the linear transformations computed using channel-wise 1x1 convolutions for the input tensors. $\boldsymbol{b}_g$ and $\boldsymbol{b}_p$ are the bias terms. $\boldsymbol{w}$ is attention coefficients computed from AG. Resampling of attention coefficients is done using nearest interpolation, which is used to match the dimensions of input features. In this way, the output of AG unit is the element-wise multiplication of input features $\boldsymbol{f}$ and attention coefficients $\boldsymbol{w}$, which identify salient input regions and prune feature responses to preserve only the relevant activations. Thus, the AG can increase prediction accuracy and improve the performance of speech enhancement task, which can be validated in the performance evaluation section.

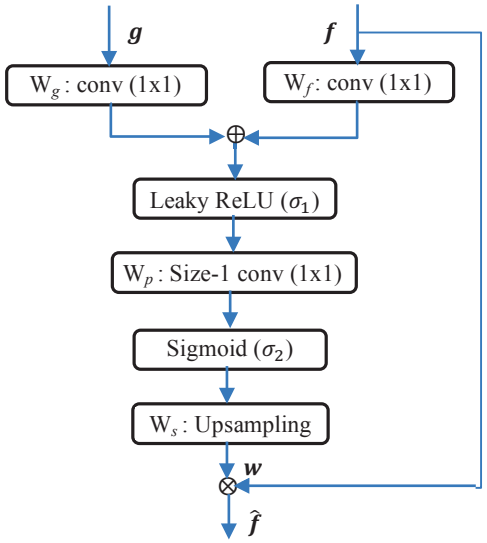

Figure 4: *Schematic of the AG Unit.*

2.4. WMAE loss

Assuming the clean speech $\boldsymbol{x}$ is contaminated by an uncorrelated additive noise $\boldsymbol{n}$, we can obtain the noisy speech signal $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n}$. By transforming $\boldsymbol{y}$ into STFT domain, we have

$$Y = X + N \quad (4)$$

where X, Y and N are the magnitude spectrogram of the clean speech, noisy speech and noise signal, respectively.

For DNN-based magnitude spectrogram estimation methods, a popular loss function is mean squared error (MSE) between clean magnitude X and estimated magnitude $\hat{X}$, which does not consider the loss of noise estimation, and thus the performance of speech estimation maybe limited. In this paper, we propose a novel noise-aware multi-task loss function, named weighted mean absolute error (WMAE) loss, in which the noise prediction loss term is also added and expected to be complementary to speech estimation. Moreover, to properly balance the contributions of each loss term and solve the scale insensitivity problem, we weighted speech and noise estimation loss term proportional to the energy of speech and noise, respectively. Therefore, the final form of the WMAE loss can be given as follows:

$$loss_{wmae}(\hat{X}, X, Y, N) = a\left|\hat{X} - X\right| + (1-a)\left|\hat{N} - N\right| \quad (5)$$

where $\hat{N} = Y - \hat{X}$ is estimated noise magnitude spectrogram and a is the weight coefficient that is the energy ratio between clean speech and noise, which is defined by

$$a = X^2/(X^2 + N^2) \quad (6)$$

3. Performance evaluation

In the test experiments, as previous speech enhancement works [17] - [22], we employed the open database presented in [23] to perform direct performance comparison. The clean speech recordings of 30 speakers were provided from the voice bank corpus [24], where 28 speakers were chosen for the training set and 2 for the test set. The noisy training set was synthesized by mixing the clean speech training set with noise from the DEMAND database [25], which includes 40 different noisy scenarios with 10 different noise conditions at signal-to-noise ratios (SNRs) of 0, 5, 10 and 15dB. The noisy test set was created using remaining 5 noise types from the DEMAND database and clean test set of 2 speakers from the Voice Bank corpus. Both speakers and noise types in the test set are unseen in the training set. The SNRs were setting at 2.5, 7.5, 12.5 and 17.5dB, respectively. Finally, the training and test set we used contain 11572 and 824 noisy-clean speech pairs, respectively.

All the training and test utterances were first downsampled from 48kHz to 16kHz. We then compute the STFT with a Blackman window of 1024 and hop length of 256 samples. Here only 512 frequency bins were taken into account and the 513th frequency bin was ignored in order to use an exact power of two that allows a simpler network. Finally, we extract a sequence of 64 frames that we feed as input and targets to the network, and the magnitude spectrograms are normalized to the range [0, 1].

The NAAGN model was trained with the Adam optimizer [26]. We set the learning rate to 0.001. We train the models with a batch size of 32. The whole model was developed in Keras [27] with Tensorflow [28] backend. The model parameters are presented in Table 1.

Table 1: *The network structure parameter of NAAGN*

Layer	Unit	Filter	Kernel	Stride
Enc.1 /Dec.1	Res. Unit	45	1x7	(1,1)
	Conv./Deconv.			(1,1)
Enc.2 /Dec.2	Res. Unit	45	7x1	(1,1)
	Conv./Deconv.			(1,1)
Enc.3 /Dec.3	Res. Unit	90	5x7	(1,1)
	Conv./Deconv.			(2,2)
Enc.4 /Dec.4	Res. Unit	90	5x7	(1,1)
	Conv./Deconv.			(1,2)
Enc.5 /Dec.5	Res. Unit	90	3x5	(1,1)
	Conv./Deconv.			(2,2)
Enc.6 /Dec.6	Res. Unit	90	3x5	(1,1)
	Conv./Deconv.			(1,2)
Enc.7 /Dec.7	Res. Unit	90	3x5	(1,1)
	Conv./Deconv.			(2,2)
Enc.8 /Dec.8	Res. Unit	90	3x5	(1,1)
	Conv./Deconv.			(1,2)
Enc.9 /Dec.9	Res. Unit	90	3x5	(1,1)
	Conv./Deconv.			(2,2)
	AG			(1,1)
Bridge	Res. Unit	180	3x5	(1,1)
	Conv./Deconv.			(1,2)

To evaluate the performance of the proposed NAAGN method, Wiener filtering (Wiener) with a priori noise SNR estimation [1], SEGAN [17], Wavenet [18], MMSE-GAN [19], Deep Feature Loss (DFL) [20], a recent hybrid model called MDPhD [21] and RSGAN-GP [22] were used as the reference methods. Meanwhile, to evaluate the contribution of AG block, the proposed ResUNet was also used as baseline approach. For the metrics to compare NAAGN and the aforementioned reference methods, the following five metrics are employed. SSNR: Segmental SNR. PESQ: Perceptual evaluation of speech quality [29]. CSIG: Mean opinion score (MOS) predictor of signal distortion [30]. CBAK: MOS predictor of background-noise intrusiveness [30]. COVL: MOS predictor of overall signal quality [30]. In addition, to show the improvement in speech intelligibility, the Short-Time Objective Intelligibility (STOI) measure [31] is applied to compare with the reference methods who ever used it in the same open database [23]. Note that all these six metrics are better if higher.

The comparison results are shown as Table 2 and Table 3. Results in Table 2 indicate that the proposed NAAGN model outperforms the reference methods with respect to all metrics by a large margin. Moreover, we also can see that the NAAGN model yield better performance than ResUNet, which proves the advantage of AG block incorporated into our method. In addition, Table 3 shows that the proposed model achieves higher speech intelligibility than reference methods.

Table 2. *Test results of objective quality*

Method	SSNR	PESQ	CSIG	CBAK	COVL
Noisy	1.68	1.97	3.35	2.44	2.63
Wiener	5.07	2.22	3.23	2.68	2.67
SEGAN	7.73	2.16	3.48	2.94	2.80
Wavenet			3.62	3.23	2.98
DFL			3.86	3.33	3.22
MMSE-GAN		2.53	3.80	3.12	3.14
MDPhD	10.22	2.70	3.85	3.39	3.27
ResUNet	10.08	2.85	4.04	3.48	3.45
NAAGN	**10.25**	**2.90**	**4.13**	**3.50**	**3.51**

Table 3. *Test results of STOI*

Method	Noisy	MMSE-GAN	RSGAN-GP	NAAGN
STOI	0.921	0.930	0.942	**0.948**

4. Conclusions

In this paper, we have proposed a NAAGN model for single channel speech enhancement, in which the residual learning, dilated convolutions and attention-gated network are incorporated into U-Net architecture. We also have designed a new noise-aware multi-task loss function, called WMAE loss, which takes the speech estimation loss and noise prediction loss into consideration simultaneously. Since the speech enhancement is treated as a sequence-to-sequence mapping problem, the NAAGN can capture long-term temporal contexts through its large receptive fields upon the input T-F representation. The attention-gated network is integrated into the proposed method, which furtherly increases the long-term contexts sensitivity and prediction accuracy with minimal computational overhead. Compared with the reference methods on the Voice Bank corpus and DEMAND database, the proposed NAAGN method shows the superior performance in all metrics, which achieves state-of-the-art performance.

5. References

[1] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Atlanta, GA, USA, vol. 2, pp. 629-632, 1996.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113-120, April 1979.

[3] C. Li and W. Liu, "A novel multi-band spectral subtraction method based on phase modification and magnitude compensation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, pp. 4760-4763, 2011.

[4] F. Deng, F. Bao, and C. C. Bao, "Speech enhancement using generalized weighted β-order spectral amplitude estimator," *Speech Communication*, vol. 59, pp. 55-68, 2014.

[5] F. Deng, C. C. Bao, and F. Bao, "A speech enhancement method by coupling speech detection and spectral amplitude estimation," *in ISCA Interspeech*, pp. 3234-3238, 2013.

[6] F. Deng, C. Bao and W. B. Kleijn, "Sparse Hidden Markov Models for Speech Enhancement in Non-Stationary Noise Environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1973-1987, Nov. 2015.

[7] K. Tan and D. L. Wang, "A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement," *in Proceedings of Interspeech*, pp. 3229-3233, 2018.

[8] J. Chen and D. L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *J. Acoust. Soc. Am.*, vol. 141, no. 6, pp. 4705-4714, 2017.

[9] K. Tan, J. Chen and D. Wang, "Gated Residual Networks with Dilated Convolutions for Monaural Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 189-198, Jan. 2019.

[10] K. Tan, J. Chen and D. Wang, "Gated Residual Networks with Dilated Convolutions for Supervised Speech Separation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, pp. 21-25, 2018.

[11] G. Naithani, T. Barker, G. Parascandolo, L. Bramsl, N. H. Pontoppidan, and T. Virtanen, "Low latency sound source separation using convolutional recurrent neural networks," in *2017 IEEE Work-shop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 71-75, 2017.

[12] D. Stoller, S. Ewert, and S. Dixon. "Wave-u-net: A multi-scale neural network for end-to-end audio source separation." In *ISMIR*, 2018.

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

[14] H.-S. Choi, J. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2019.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 770-778, 2016.

[16] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz et al., "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[17] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network." In *Proc. Interspeech 2017*, pp. 3642-3646, 2017.

[18] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising." *In IEEE ICASSP 2018*, pp. 5069-5073. 2018.

[19] F. G. Germain, Q. Chen and V. Koltun, "Speech denoising with deep feature losses." *arXiv preprint arXiv:1806.*10522, 2018.

[20] M.H. Soni, N. Shah and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network." *In IEEE ICASSP 2018*, pp. 5039-5043, 2018.

[21] J.-H. Kim, J. Yoo, S. Chun, A. Kim, and J.-W Ha, "Multi-domain processing via hybrid denoising networks for speech enhancement." *arXiv preprint arXiv:1812.08914,* 2018.

[22] D. Baby and S. Verhulst, "Sergan: Speech Enhancement Using Relativistic Generative Adversarial Networks with Gradient Penalty," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, pp. 106-110, 2019.

[23] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise robust text-to-speech," in *ISCA-SSW*, pp. 146-152, 2016.

[24] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *IEEE-OCOCOSDA/CASLRE* pp. 1-4, Nov 2013.

[25] 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*, vol. 19, no. 1, p. 035081, 2013.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[27] F. Chollet et al., "Keras," https://keras.io, 2015.

[28] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/.

[29] ITU, Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs ITU-T Rec. P.862, 2000.

[30] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement." I*EEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp: 229-238, 2007.

[31] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE-TASLP*, vol. 19, no. 7, pp. 2125-2136, Sep 2011.