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

Noisy-reverberant Speech Enhancement Using DenseUNet with Time-frequency Attention

Yan Zhao and DeLiang Wang

Department of Computer Science and Engineering
& Center for Cognitive and Brain Sciences, The Ohio State University, USA
zhao.836@osu.edu, dwang@cse.ohio-state.edu

Abstract

Background noise and room reverberation are two major distortions to the speech signal in real-world environments. Each of them degrades speech intelligibility and quality, and their combined effects are especially detrimental. In this paper, we propose a DenseUNet based model for noisy-reverberant speech enhancement, where a novel time-frequency (T-F) attention mechanism is introduced to aggregate contextual information among different T-F units efficiently and a channelwise attention is developed to merge sources of information among different feature maps. In addition, we introduce a normalization-activation strategy to alleviate the performance drop for small batch training. Systematic evaluations demonstrate that the proposed algorithm substantially improves objective speech intelligibility and quality in various noisy-reverberant conditions, and outperforms other related methods.

Index Terms: speech denoising, speech dereverberation, complex ratio mask, DenseUNet, attention

1. Introduction

In daily listening environments, speech is inevitably corrupted by background noise. Besides additive noises, reverberation caused by the attenuated and delayed reflections of sound waves in a room is another major distortion that we face everyday. These distortions together degrade both speech intelligibility and quality, especially when the signal-to-noise ratio (SNR) is low [1, 2]. Furthermore, many speech processing tasks such as automatic speech recognition (ASR) and speaker identification (SID) become more difficult under these adverse noisy-reverberant conditions [3, 4].

For noisy-reverberant speech enhancement, Han *et al.* [5] proposed to utlize deep neural networks (DNNs) to learn a nonlinear mapping from the log magnitude spectrum of noisy-reverberant speech to that of clean-anechoic speech. Considering the different natures of background noise and room reverberation, Zhao *et al.* [6] employed a two-stage strategy to enhance noisy-reverberant speech, where noise and reverberation were removed in two separate stages, respectively. The two-stage system was jointly optimized with a time loss during training. Similar to [5], during testing, the time-domain signals were resynthesized using the Griffin-Lim phase enhancement algorithm [7]. Ribas *et al.* [8] proposed a wide residual network (WRN) based model to perform enhancement, which leverages the residual connections in a very deep architecture. Although good performance was obtained, these methods focus on performing enhancement in the magnitude domain, while leaving phase enhancement to a post processing stage or simply using the corrupted phase. Recent studies [9, 10] have shown substantial improvements of performing enhancement in the complex domain. In addition, speech separation studies [11, 12] were conducted on time-domain signals in an end-to-end fashion. However, the short-time Fourier transform (STFT) representation may be more stable than the representation directly learned from the waveform signals. This paper develops a noisy-reverberant speech enhancement algorithm in the complex domain.

From the perspective of DNN models, we adopt a DenseUNet architecture as the backbone, which is a combination of the architecture of UNet [13] and DenseNet [14]. It is worth noting that DenseUNet based models have been successfully employed in several speech processing tasks, such as speaker separation [15] and speech enhancement [16]. In this study, we further improve the DenseUNet model for noisy-reverberant speech enhancement. Specifically, we propose a novel time-frequency (T-F) attention mechanism to integrate global information among different T-F units and design a channelwise attention mechanism to merge feature maps according to their importance. Attention-based models have been studied previously for speech enhancement [17, 18, 16].

The rest of this paper is organized as follows. We describe the proposed algorithm in the next section. The experimental setup and evaluations are presented in Section 3 and Section 4. Section 5 concludes this paper.

2. Algorithm description

2.1. Problem formulation

Let $s(t)$, $h(t)$ and $n(t)$ denote clean speech, room impulse response (RIR) function, and background noise, respectively. The noisy-reverberant speech $y(t)$ can be written as

$$y(t) = s(t) * h(t) + n(t) = x(t) + r(t) + n(t) \quad (1)$$

where $*$ stands for the convolution operator; $x(t)$ and $r(t)$ denote anechoic speech (direct sound) and its reverberation, respectively. Our objective is to recover $x(t)$ from the observed $y(t)$.

Given noisy-reverberant speech sampled at 16 kHz, features are extracted by framing the signal using the 32-ms Hamming window with a 8-ms window shift, and then applying a 512-point fast Fourier transform (FFT) to each frame. This results in 257 frequency bins. Supposing the number of frames in the utterance is N, we use Y to denote the extracted features in the frequency domain, which is a $N \times 257$ matrix with complex values. Similarly, let X denote the T-F representation of $x(t)$.

This study was supported in part by an NIH grant (R01 DC012048) and the Ohio Supercomputer Center.

http://dx.doi.org/10.21437/Interspeech.2020-2952

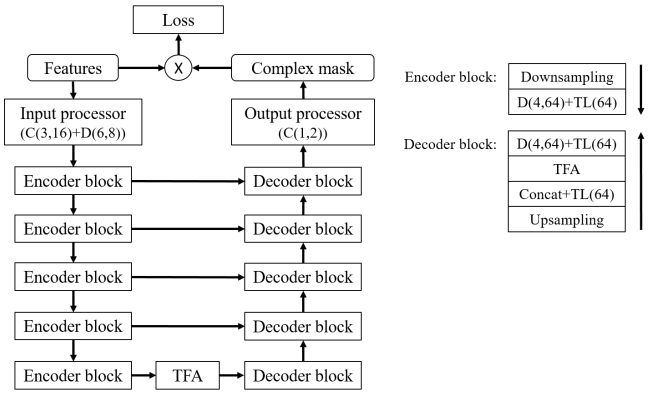


Figure 1: *System diagram of the proposed noisy-reverberant speech enhancement model. "C(·, ·)" denotes a basic convolutional block, "D(·, ·)" denotes a dense block, "TL(·)" denotes a transition layer, "Concat" denotes concatenating feature maps, and "TFA" denotes the proposed T-F attention. The data flow in encoder blocks is top-down, in decoder blocks it is bottom-up.*

2.2. System overview

Fig. 1 shows the diagram of the proposed system, which takes the real and imaginary components of Y as two feature maps. We pass the features to an input processor, where the number of channels gradually increases to 64 by employing a convolutional block and a dense block. In this study, we employ 5 dense blocks for the encoder and 5 for the decoder. Max pooling is used to perform downsampling during encoding, and transposed convolution is utilized for upsampling during decoding. The output of the DenseUNet is passed to an output processor, which includes a convolutional block to reduce the number of channels to 2. These two feature maps are interpreted as the real and imaginary components of an estimated complex ratio mask (cRM). When applied to the complex spectrum of noisy-reverberant speech, it provides an estimate of enhanced complex spectrum $\hat{X}$, namely,

$$\hat{X} = cRM \otimes Y \tag{2}$$

Then, we compute the loss as follows,

$$\mathcal{L} = \|X_r - \hat{X}_r\|_1 + \|X_i - \hat{X}_i\|_1 + \|\|X\|_2 - \|\hat{X}\|_2\|_1 \tag{3}$$

where $\|\cdot\|_1$ denotes the $\mathcal{L}_1$ norm and $\|\cdot\|_2$ denotes the $\mathcal{L}_2$ norm; subscript r and i denote the real and imaginary components of the complex spectrum, respectively. By incorporating a magnitude loss as part of the loss function, better PESQ [19] results can be expected [20].

Next we explain several terms and blocks used in the system description. Note that the terms *feature maps* and *channels* are used interchangeably in the following descriptions.

A basic convolutional block is denoted by C(*kernel_size*, *out_channel*), which consists of a filter response normalization (FRN) with the Thresholded Linear Unit (TLU) activation (we use the term FRN layer to denote a FRN with the TLU) [21] and a depthwise separable convolutional layer [22]. The kernel size of the convolutional layer is given by *kernel_size*. The number of output channels is given by *out_channel*. Due to the design of UNet and the heavy use of dense blocks, the GPU memory consumption for training becomes very large. The introduction of the attention makes the memory situation even worse. So we have to use a small batch size (4 in our experiments). However, when the batch size becomes too small, the performance of batch normalization deteriorates rapidly. We adopt FRN layers since our study shows better performance using FRN layers than other normalization techniques with small batch size. On the other hand, compared with standard convolution, depthwise separable convolution provides good performance using much fewer parameters [22, 23].

Within a dense block, at layer l, the feature maps of all preceding layers, $\boldsymbol{z}_0$, $\boldsymbol{z}_1$, ..., $\boldsymbol{z}_{l-1}$, are simply concatenated and fed as input:

$$\boldsymbol{z}_l = H_l([\boldsymbol{z_0}, \boldsymbol{z_1}, ..., \boldsymbol{z}_{l-1}]) \tag{4}$$

where H_l denotes the convolutional block at layer l. If H_l produces k feature maps and the number of input channels of the dense block is k_0, the number of output channels at layer l is $k_0 + k \times l$. Symbol k is also called the growth rate of the dense block. In our system, a dense block with L layers is denoted by D(L, *growth_rate*). In order to control the number of channels to some level, after each dense block, we add a transition layer (TL) to reduce the channels. Let TL(*out_channel*) denote the transition layer, which can reduce the number of input channels to the given *out_channel*.

2.3. Time-frequency attention

For speech processing tasks, the contextual information plays a key role. To leverage such information, a general way is to employ an attention mechanism [24]. Note that the contextual information is not limited to the time dimension. For speech, the frequency dimension should be also taken into account.

Following the terminology in [24], we first describe how to compute attention and produce new representations in general. An attention module takes queries (Q) and key-value ($K - V$) pairs as the input. Mathematically, Q, K and V are matrices and considered as a set of vectors (queries, keys or values) with each row vector as the element. Inside the attention module, they are first linearly projected to Q', K' and V', respectively. Given a query $\boldsymbol{q'}$ from Q', a weight distribution on the keys set is computed by the similarities between the query ($\boldsymbol{q'}$) and keys (K'). Then, by making a weighted sum of the values (V') with the computed weight distribution, a dynamic representation is obtained to capture more relevant information in the key-value pairs. Typically a SoftMax function is applied to these weights, making the sum of the weights to be 1. There are different choices for the similarity computation, like inner product or using a small neural network. In our study, we use a scaled dot product as the similarity function. Therefore, the attention is computed by

$$Attention(Q', K', V') = SoftMax(\frac{Q'K'^T}{\sqrt{d_{\boldsymbol{k'}}}})V' \tag{5}$$

where $d_{\boldsymbol{k'}}$ is the dimension of vector in keys set. If queries, keys, and values are from the same vector set, it becomes self attention. The proposed T-F attention is based on self attention. After applying attention weights, the weighted sum is projected back to the original space and added to the original representation to produce a new representation.

The intermediate T-F representations (feature maps) learned by the DenseUNet are tensors with the dimension $T \times F \times D$ ($D = 64$ in our experiments). Since we perform downsampling or upsampling on both time and frequency dimensions during encoding and decoding, both T and F change after encoder blocks or decoder blocks. If we consider each T-F unit as an element, such a representation can be viewed as a

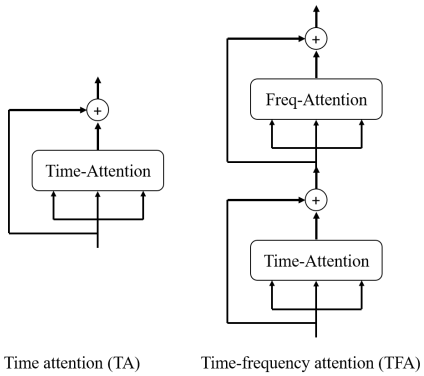
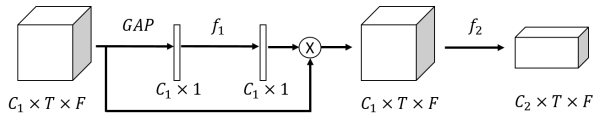


Figure 2: *Diagrams of the proposed time attention (TA) and time-frequency attention (TFA).*

vector set, in which each vector's dimension is D and the number of vectors is $T \times F$. To explore the global information among different T-F units, one natural idea is to take this vector set as the input for an attention module. However, for speech processing applications, this idea becomes almost impossible to implement in practice. With the vector set size to be $T \times F$, the size of attention weight matrix in the attention module is $(T \times F) \times (T \times F)$, which would be very large.

To address this dimension explosion problem in directly employing the attention mechanism, we propose to factorize the T-F attention into time attention (TA) and frequency attention (FA). Fig. 2 shows the diagram of the proposed T-F attention (TFA). By doing this factorization, we reduce the large attention weight matrix to two much smaller ones, $T \times T$ and $F \times F$. More importantly, it does not affect the aggregation of information among different T-F units. With the TFA, the information is first integrated to an intermediate T-F unit through the time path (by employing TA), and then integrated to the target one through the frequency path (by employing FA). The order of TA and FA is not important. In our study, we find that the two orders (TA+FA or FA+TA) perform similarly.

2.4. Improved transition layer



GAP: Global average pooling

f_1: 1-D Conv -> ReLU() -> 1-D Conv -> Sigmoid()

f_2: FRN -> 1x1 Conv

Figure 3: *A schematic diagram of the proposed transition layer. The number of channels is reduced from C_1 to C_2*

Transition layers are used in our model to reduce the number of channels after dense blocks or the concatenation of encoding features and decoding features. One simple choice is to use a 1×1 convolutional block, which treats all the channels with the same importance when merging them. Obviously, different feature maps have different contributions. To leverage such channel information, we propose to add a channelwise attention before the 1×1 convolutional block to build the transition layer. Fig. 3 shows the diagram of the proposed transition layer. Different from [25], two 1-D convolutional layers instead of fully-connected layers are used in f_1 to reduce the model complexity and the number of introduced parameters. Specifically, we first apply global average pooling to obtain the global information of each channel, i.e. mean statistics of each channel. Then an excitation function f_1 is applied to get a weight for each channel. We use the obtained weight vector to recalibrate the original features. Finally, the number of channels is reduced by using f_2 (a FRN layer plus a 1×1 convolutional block).

3. Experimental setup

3.1. Datasets

We generate noisy-reverberant data using the WSJ0 corpus [26] as target speech. From the corpus, 7138, 410 and 330 clean utterances are selected to produce training data, validation data and test data, respectively. There are 83 speakers in the training data; 10 speakers in the validation data; and 12 speakers in the test data. All the speakers in the test data are unseen during training. Three reverberant rooms, from small to large, are simulated. We place the microphone in a fixed position in each room, and randomly select the position of speaker with two distances from the microphone, namely, near (0.5 m) and far (2 m). Reverberation time is selected from 0.3 s to 1.0 s, with a 0.1 s increment. We employ an RIR generator[1] to generate 6 RIRs for each utterance in the training and validation data using randomly chosen reverberant configurations. To investigate the generalization ability of the proposed systems to different reverberant conditions, different test sets are generated. See Section 4 for details.

We use different noises for training, validation and testing. For training, we utilize 10,000 noises from a sound effect library[2]. The total duration of the noises is about 126 hours. Each reverberant utterance is mixed with 4 random noise cuts at a random SNR chosen from -6 dB to 0 dB with 1 dB increment. To further augment the noise data, one third of the noise cuts are not sampled from the 10k noise but correspond to mixed noises. Specifically, we first randomly sample 4 different cuts and then mix them together to produce a mixed noise cut. The use of mixed noises is to further increase the noise variety. Just like a noise cut from the orignal set of noises, a mixed noise cut is added to a reverberant utterance at a chosen SNR. There are 42828 (reverberant speech) $\times$ 4 (noise cuts) = 171312 noisy-reverberant utterances for training. For the validation data, three noises from the DEMAND corpus [27] are selected. The first channel signal of the corpus is used for data generation. These selected noises were recorded in a busy subway station, an office cafeteria, and a university restaurant. Each reverberant utterance in the validation set is mixed with a random cut from these three noises at a randomly chosen SNR from -6 dB to 6 dB with 1 dB increment. We have 2460 noisy-reverberant utterances for the validation set.

To simulate realistic noisy-reverberant room environments, two different room noises from the DEMAND corpus are selected to generate test data. These noises were recorded in a living room and an office, respectively. Three SNRs, -6 dB, 0 dB and 6 dB, are chosen to perform evaluation. The reverberant speech is taken as the reference signal when computing the SNR. Since the objective is to recover the clean-anechoic speech from its noisy-reverberant observation, the ac-

[1]Available at `https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator`.

[2]Available at `https://www.sound-ideas.com/`.

tual SNR level (taking clean-anechoic speech as the signal) becomes much lower, which makes it very challenging to enhance.

3.2. Comparison systems

We take the DenseUNet using FRN layers and the proposed transition layers as a baseline system for comparison, and denote it as **"DenseUNet"**. The architecture is similar to that used in deep CASA [15]. In order to investigate the function of our proposed attention mechanism, we denote the proposed system with the time attention as **"DenseUNet+TA"**, with the T-F attention as **"DenseUNet+TFA"**.

We also employ an improved two-stage model [6] as another baseline for comparison. Compared with the original method in [6], there are several differences. Firstly, we replace the feedforward neural networks with bidirectional long short-term memory networks (BLSTMs). Secondly, we use $\mathcal{L}_1$ loss to perform optimization instead of $\mathcal{L}_2$ loss. Thirdly, for the dereverberation stage, we directly predict the log magnitude spectrum of the clean-anechoic speech. For fair comparisons, the log magnitude spectrum of the noisy-reverberant speech is used as input features instead of using the complementary features. In our experiments, for each stage, a two layer BLSTM network with 1024 hidden units is used, with 512 units assigned to each direction. We denote this system as **"BLSTM-2S"**.

4. Evaluations and comparisons

In this study, STOI [28] is employed to evaluate speech intelligibility, and PESQ [19] is employed to evaluate speech quality. The value range of STOI is typically [0, 1], and for PESQ, it is [-0.5, 4.5].

4.1. Evaluations on untrained RIRs

To study whether the trained models generalize well to other rooms, we generate a new set of RIRs. Specifically, we simulate a room with size 10 m $\times$ 7 m $\times$ 3 m. Three reverberation times (0.3 s, 0.6 s and 0.9 s) are investigated, and these T_{60} values have been used during training. The microphone-speaker distance is set to 2 m. All the RIRs and sentences used for testing are unseen during training and validation. Two new noises are added at three SNRs (-6 dB, 0 dB and 6 dB) as described in Section 3.1. For convenience, when performing evaluation, we take average across these three reverberation times at each SNR.

Table 1: *Average STOI and PESQ scores at untrained RIRs for LIVING-ROOM noise. Boldface indicates the best performance.*

	STOI (in %)				**PESQ**			
SNR (dB)	**-6**	**0**	**6**	**Avg.**	**-6**	**0**	**6**	**Avg.**
mixture	62.10	68.25	72.13	67.49	1.53	1.76	1.90	1.73
BLSTM-2S	82.61	87.12	89.02	86.25	2.30	2.55	2.66	2.50
DenseUNet	82.66	87.93	90.33	86.97	2.25	2.56	2.73	2.51
DenseUNet+TA	85.68	90.64	92.93	89.75	2.45	2.75	2.93	2.71
DenseUNet+TFA	**86.82**	**91.28**	**93.31**	**90.47**	**2.52**	**2.84**	**3.02**	**2.79**

The evaluation results are shown in Table 1 and Table 2. Similar performance trend can be observed with different noise types, so we take the living-room noise condition as an example. In Table 1, all the enhancement algorithms improve STOI and PESQ substantially, indicating improvements on both speech intelligibility and quality. On average, adding the time attention improves the baseline DenseUNet model by 2.78% in STOI and by 0.20 in PESQ. This demonstrates that aggregating the information across the time dimension boosts the performance significantly. As we have studied in [29], the contextual information in the time dimension is important for speech dereverberation. Considering the reverberation components in noisy-reverberant speech, such improvements are to be expected. Moreover, with the frequency attention, the performance is improved further. The proposed DenseUNet+TFA model outperforms the DenseUNet+TA model by 0.72% in STOI and by 0.08 in PESQ.

Table 2: *Average STOI and PESQ scores at untrained RIRs for OFFICE-ROOM noise.*

	STOI (in %)				**PESQ**			
SNR (dB)	**-6**	**0**	**6**	**Avg.**	**-6**	**0**	**6**	**Avg.**
mixture	73.25	74.83	75.39	74.49	1.93	1.99	2.02	1.98
BLSTM-2S	89.06	89.79	90.09	89.65	2.65	2.72	2.75	2.71
DenseUNet	89.42	91.02	91.79	90.74	2.68	2.80	2.86	2.78
DenseUNet+TA	92.45	93.86	94.52	93.61	2.90	3.03	3.10	3.01
DenseUNet+TFA	**92.63**	**94.05**	**94.72**	**93.80**	**2.98**	**3.11**	**3.18**	**3.09**

In addition, the two-stage baseline system shows comparable performance with the DenseUNet model. For most noisy-reverberant conditions, the DenseUNet model only slightly outperforms the BLSTM-2S model.

4.2. Evaluations on recorded RIRs

Previously, we employ simulated RIRs. Now we evaluate the proposed algorithms with recorded RIRs. Two RIRs from the Aachen Impulse Response (AIR) database [30] are selected, and are resampled to 16 kHz. They were recorded in a living room and an office, and reverberation times are about 0.70 s and 0.37 s, respectively. The living-room noise and the office-room noise from the DEMAND corpus is added to the corresponding room to produce noisy-reverberant speech at -6 dB SNR.

Table 3: *Average STOI and PESQ scores with recorded RIRs at -6 dB SNR.*

	STOI (in %)		**PESQ**	
room	**living**	**office**	**living**	**office**
mixture	68.05	82.78	1.58	2.19
BLSTM-2S	85.86	92.14	2.41	2.85
DenseUNet	85.96	92.75	2.38	2.93
DenseUNet+TA	87.60	**93.86**	2.50	3.01
DenseUNet+TFA	**88.31**	93.67	**2.56**	**3.06**

Table 3 presents the evaluation results. All the enhancement systems improve the objective speech intelligibility and quality of noisy-reverberant speech substantially. This demonstrates that models trained with simulated RIRs are able to generalize to real recorded RIRs well. The proposed DenseUNet+TFA model performs the best under most conditions.

5. Conclusion

Motivated by the need to aggregate contextual information among different T-F units, we have proposed a T-F attention mechanism to improve DenseUNet based noisy-reverberant speech enhancement. In addition, the proposed system performs the enhancement in the complex domain by implicitly estimating a complex ratio mask. In other words, the magnitude spectrum and phase spectrum are jointly enhanced. Systematic evaluations demonstrate that our proposed system is able to remove both background noise and room reverberation effectively, and outperforms previous two-stage models for noisy-reverberant speech enhancement.

6. References

[1] A. K. Nábělek, "Communication in noisy and reverberant environments," *Acoustical factors affecting hearing aid performance*, pp. 15–28, 1993.

[2] B. Edwards, "The future of hearing aid technology," *Trends in amplification*, vol. 11, pp. 31–45, 2007.

[3] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, pp. 745–777, 2014.

[4] K. A. Al-Karawi, A. H. Al-Noori, F. F. Li, and T. Ritchings, "Automatic speaker recognition system in adverse conditions - implication of noise and reverberation on system performance," *International Journal of Information and Electronics Engineering*, vol. 5, pp. 423–427, 2015.

[5] K. Han, Y. Wang, D. L. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 982–992, 2015.

[6] Y. Zhao, Z.-Q. Wang, and D. L. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM Trans. Audio. Speech Lang. Proc.*, vol. 27, pp. 53–62, 2019.

[7] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 236–243, 1984.

[8] D. Ribas, J. Llombart, A. Miguel, and L. Vicente, "Deep speech enhancement for reverberated and noisy signals using wide residual networks," *arXiv preprint arXiv:1901.00660*, 2019.

[9] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, pp. 483–492, 2016.

[10] K. Tan and D. L. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 380–390, 2020.

[11] A. Pandey and D. L. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 1179–1188, 2019.

[12] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 1256–1266, 2019.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*. Springer, 2015, pp. 234–241.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, 2017, pp. 4700–4708.

[15] Y. Liu and D. L. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 2092–2102, 2019.

[16] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-Attention Dense U-Net for multichannel speech enhancement," *arXiv preprint arXiv:2001.11542*, 2020.

[17] X. Hao, C. Shan, Y. Xu, S. Sun, and L. Xie, "An attention-based neural network approach for single channel speech enhancement," in *Proc. ICASSP*, 2019, pp. 6895–6899.

[18] C.-F. Liao, Y. Tsao, X. Lu, and H. Kawai, "Incorporating symbolic sequential modeling for speech enhancement," in *Proc. INTERSPEECH*, 2019, pp. 2733–2737.

[19] 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*, vol. 2, 2001, pp. 749–752.

[20] Z.-Q. Wang and D. L. Wang, "Deep learning based target cancellation for speech dereverberation," *IEEE/ACM Trans. Audio. Speech Lang. Proc.*, vol. 28, pp. 941–950, 2020.

[21] S. Singh and S. Krishnan, "Filter response normalization layer: eliminating batch dependence in the training of deep neural networks," *arXiv preprint arXiv:1911.09737*, 2019.

[22] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proc. CVPR*, 2017, pp. 1251–1258.

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[26] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the workshop on speech and natural language*, 1992, pp. 357–362.

[27] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoust. Soc. Am.*, vol. 133, pp. 3591–3591, 2013.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 2125–2136, 2011.

[29] Y. Zhao, D. L. Wang, B. Xu, and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE/ACM Trans. Audio. Speech Lang. Proc.*, vol. 28, pp. 1598–1607, 2020.

[30] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. ICDSP*, 2009, pp. 1–5.