DRC-NET: DENSELY CONNECTED RECURRENT CONVOLUTIONAL NEURAL NETWORK FOR SPEECH DEREVERBERATION

Jinjiang Liu, Xueliang Zhang

College of Computer Science, Inner Mongolia University, China
jetliu1994@foxmail.com, cszxl@imu.edu.cn

ABSTRACT

Under our previous work on frequency bin-wise independent processing, a dramatic reduction of the computational complexity for recurrent neural networks (RNN) is achieved. So that a massive deployment of RNN in time dimension is realized in this paper, by using the channel-wise long short-term memory neural network. Based on this approach, the processing of RNN on frequency dimension and time dimension in the time-frequency domain are unified. This allows us to combine convolutional neural network (CNN) and RNN as a basic neural operator, which finally leads to the Densely Connected Recurrent Convolutional Neural Network (DRC-NET). The DRC-NET sufficiently exploits the infinite response of RNN, and the finite response of CNN. Its balanced response characteristics significantly improve the system performance. Experimental result shows that both non-causal and causal version of DRC-NET outperforms the state-of-the-art (STOA) model for speech dereverberation task.

Index Terms— speech dereverberation, microphone array processing, convolutional recurrent neural network, deep learning

1. INTRODUCTION

Reverberations are widely experienced in daily life, which is caused by the multi-path transmission of sonic wave in an enclosed space. The reflections and energy loss during the acoustic transmission make it a mixture of infinite numbers of attenuated speech copies. Thus, a strong correlation exists between speech and its reverberation which is difficult to eliminate. Even the modest speech reverberation approach could lead to a detrimental impact on speech intelligibility. For example, in a heavily reverberant room, the speech captured by a far-field microphone sounds blurred and confusing in teleconferencing, people with hearing impairment could not track the discussions.

The dereverberation is a process of removing reverberations or late reverberation components from the observed mixture, usually by signal processing or data-driven deep learning. In practice, the reverberation can be mathematically modeled as the convolution process between the source signal and the acoustic impulse response (AIR). Harmonicity-based dereverberation (HERB) [1] first estimates a harmonic direct signal by adapting a time-varying harmonic comb filter, and then forms a time-invariant dereverberation filter by the observed signal and the estimated harmonic direct arrive signal. It performs well but with high latency to calculate an accurate inverse filter. The well known weighted prediction error (WPE) [2][3] is a widely used dereverberation algorithm, which can work for both single-channel and multi-channel conditions. The WPE models the dereverberation as an auto-regressive process. As a parameter to calculate the dereverberation filter, the short-time power spectral density (PSD) of the speech is estimated iteratively. Due to the linear prediction, WPE has no non-linear distortion to signal phase. For multi-channel case, spatial directivity of beamformer or channel correlations of the microphone array [4] can be utilized to effectively suppress reverberations by a spatial beam pattern [5], which can also be combined with WPE [6][7].

All those conventional algorithms are formulated under concrete and complete theoretical fundamentals. But the reverberations in real world are too complex to be fit mathematically by handcraft model. Instead, the data-driven deep neural network (DNN) model with powerful non-linear fitting ability shows its promising performance recently [8]. Some of researches, such as in [9], combine deep learning methods with conventional methods, mostly by using DNN to robustly estimate the critical parameters in those conventional methods. These combined methods usually perform well with lower computational complexity but are limited by the upper bound of the conventional part. Currently, the end-to-end neural network for speech signal processing becomes popular with impressive performance. There are mainly two ways to promote the neural network performance, one is to properly design the learning target and the output constrain [10][11][12]. The other one is to design topology of the model or neural operator which could effectively fit the problems [13][14][15]. Convolutional Recurrent Networks (CRN) are representative models [16][17][18] for speech enhancement, which could learn speech pattern effectively by combining CNN and RNN together.

The dense convolutional neural network (DenseNET) [19] has been utilized recently for speech dereverberation [20][21]. A substantial improvement was reported which makes it the SOTA algorithm. The DenseNET is a very deep model which has a total of 56 layers. Its dominant part is the 54-layer CNN, and only 2-layer LSTM are embedded in the middle of the U-Net structure. However, within such a structure, the LSTM might not be fully utilized. In fact, the mechanism of the CNN and RNN are quite different which can be viewed as finite and infinite impulse response filter, respectively. It motivates us that a sufficient and complete combination of these two mechanisms is reasonable and promising. In our previous work for dual-channel speech enhancement [15], By using the channel-wise LSTM, we explore the independence of frequency bin-wise processing, which improves the performance and significantly reduces the computational complexity. In the same way, RNN can be equipped in DRC-NET on a large scale.

Our study makes three main contributions in this work. First, we introduce channel-wise LSTM and unify its usage in time and frequency dimension. Second, we propose a recurrent convolutional (RC) unit which combines channel-wise LSTM and CNN. Third, we utilize the RC unit to construct a densely connected recurrent convolutional model for speech dereverberation, which dramatically improves the performance compared to the SOTA algorithm.

This research is supported by the National Natural Science Foundation of China (No. 61876214).

2. ALGORITHM

2.1. Reverberation model and objectives

For m-channel microphone setup, in STFT domain, the signal captured by the microphone array can be modeled as follows

$$X_{n,k}^{m} = \sum_{l=0}^{L_h-1} (H_{l,k}^{m})^{*} S_{n-l,k} + N_{n,k}^{m}, \tag{1}$$

where, n and k represent the time and frequency bin index. $X_{n,k}^{m}$, $H_{l,k}^{m}$, $S_{n-l,k}$, and $N_{n,k}^{m}$ are the observed mixture, AIR, target speech, and noise signal, respectively. The goal of this work is to recover the early speech component of the first channel from the observed reverberant mixtures $X_{n,k}^{m}$.

2.2. Complex mapping for reverberation suppression

Domain selection, model design, and model constrain are very important for deep learning based algorithms. Although time-domain models [14][13] are the SOTA approach for speech enhancement, our algorithm works on T-F domain for the following considerations. First, the time duration of the reverberations could be very long in time domain, usually varying from thousand to tens of thousands of samples. So, it is difficult to trace such a long-term pattern. But in T-F domain, the duration could be significantly shortened by folding it through the frame shift operation of STFT. Second, the phase delay and energy decay for different frequencies are naturally independent. While in time domain those are all blended. So the dereverberation in T-F domain should be easier. Third, a proper model design for a specific transform domain is important. In current works [22] [23], RNNs are not fully utilized because they only appear after CNN encoder and do not participate in the whole process of encoding and decoding.

Directly mapping complex spectrum has been proofed to be a neat and effective way to estimate target signal for neural networks. Many recent studies [12][15][21] show that a direct constrain on amplitude is meaningful for the perceptual quality, and the L1-norm is a better measure to form a correct phase estimation. In this paper, a complex mapping with amplitude and phase constrained by L1-norm is implemented exactly as it was implemented in [21] as follows

$$\begin{aligned} \mathcal{L} = &\left\| |\hat{S}| - |S| \right\|_1 + \\ &\left\| Real(\hat{S}) - Real(S) \right\|_1 + \left\| Imag(\hat{S}) - Imag(S) \right\|_1 \end{aligned} \tag{2}$$

where, S and $\hat{S}$ are the target and the estimated spectra, $Real(\cdot)$ and $Imag(\cdot)$ are the operation to extract the real and imaginary components for a complex spectrum, and $\|\cdot\|_1$ is the $L1$ norm.

2.3. Key mechanisms for DRC-NET

2.3.1. Massively embedded channel-wise LSTM

In CRN model, LSTM plays an important role in modeling the temporal patterns of speech spectral. For the powerful DenseNET [21], the effectiveness of LSTM is limited, which is embedded in the bottleneck layer of the U-Net structure. Even with the dense connection, the gradients are still hard to be sufficiently transmitted to such a deep structure.

The channel-wise LSTM is a novel way of implementing the LSTM model in the T-F domain that applies a shared LSTM on each frequency bin. Technically, for a T-F domain feature map in a shape of $[\mathbf{B}atch, \mathbf{C}hannel, \mathbf{F}requency, \mathbf{T}ime]$, it could be conveniently realized by transpose and merge frequency dimension into the batch dimension to the shape of $[\mathbf{B} \times \mathbf{F}, \mathbf{T}, \mathbf{C}]$ as the input feature of LSTM, and reshape the output of LSTM back to $[\mathbf{B}, \mathbf{C}, \mathbf{F}, \mathbf{T}]$.

Our previous study [15] showed that the frequency bin-wise processing of LSTM could effectively separate speech components from noisy features. Especially for the multi-channel microphone task, the channel-wise LSTM can effectively use the spatial cues in each frequency. The proposed DRC-NET in this work follows the similar idea and is further extended. The processing of LSTM long time or frequency dimension now is unified by utilizing channel-wise LSTM, just like CNN could be either applied along the time or along the frequency axis.

2.3.2. Recurrent convolutional (RC) unit

RC unit is a basic unit of dense block in DRC-NET, which consists of a LSTM layer and a convolutional layer. As we know, the characteristics of the time-frequency response of FIR and IIR are very different. In an FIR system, frequencies are shifted in time by the same amount, so its phase response is linear with no feedback to the system. This leads to a consistent system stability [24]. While, the IIR system could achieve better amplitude response control by constantly accumulating its historical response. The very nature of those two systems causes their significantly different characteristics, the output of the FIR system is linear combinations of finite inputs, while the IIR accumulates infinite response to form current response, it is easy to control amplitude, but phase could be blurred by it infinitely accumulated residual components. In the view of information theory, with the same entropy capacity, the FIR system realizes accurate phase response and IIR system realizes better magnitude response control.

In neural network design, CNN and RNN are currently two fundamental neural operators, and conceptually can be viewed as neural FIR and neural IIR. By the associative law of convolution, stacked CNN layers with the small kernel can mathematically reform an equivalent long FIR filter, and the activation function provides the non-linearity for both amplitude and phase. Different from the unstable phase response of IIR, this FIR non-linearity for phase is desired for it can be utilized for nonlinear phase modeling. LSTM can be viewed as an IIR system. We believe that IIR and FIR systems with different response characteristics could be complementary, and their combination would lead to a promising performance.

Technically, we form an RC(T/F) unit by stacking a channel-wise LSTM in time (T) or frequency (F) and a convolutional layer, as shown in Fig. 1. When LSTM being applied on frequency dimension RC(F) unit, we use bidirectional LSTM. If LSTM is applied in time dimension RC(T), the LSTM is unidirectional for causal system and is bidirectional for non-causal system. The kernels (frequency x time) of convolution are 3x1 and 3x3 for causal and non-causal system, respectively. By the proposed channel-wise LSTM, we unify the processing of LSTM on the time and frequency axis, and control the computational complexity and model size in a reasonable level.

2.3.3. Dense block with RC unit

In previous studies [19][21], the dense connection are utilized to enhance the performance of CNN. In this work, as shown in Fig. 1, we combines the powerful RC unit with dense connectivity as densely connected recurrent convolutional (DRC) block. There are 4 RC unit alternatively processing on frequency and time axis. So, the DRC block is a deep structure with a total of 8 layers.

As we know, the depth of neural networks profoundly affects the performance, and the dense connection has been proved to be an

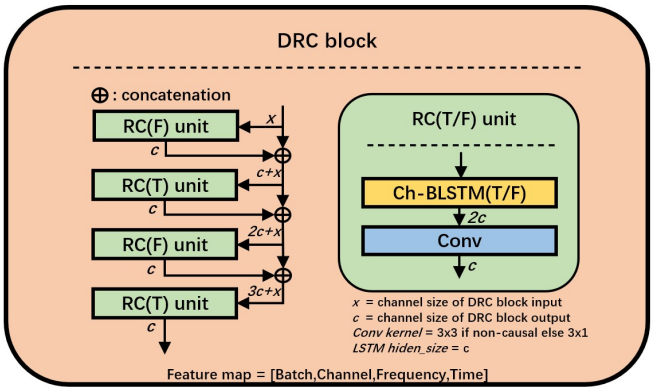

Fig. 1. The architecture of dense RC block and RC unit

effective to making network deep and avoiding gradient vanishing. A theoretical work in the field of dynamic systems illustrated why the dense connection works [25]. For a dual-branch neural network layer, it can be expressed as follows

$$z_{l+1} = G(h(z_l) + \mathscr{F}(z_l, W_l)) \tag{3}$$

where z_l and z_{l+1} are the input and outputs of the system, $h(.)$ and $G(.)$ are the nonlinear maps, and $\mathscr{F}$ is a small perturbation when $h(.)$ and $G(.)$ close to the identity map, the gradient of the right-hand side would close an identity map, as follows

$$z_{l+1} = z_l + \mathscr{F}(z_l, W_l) \tag{4}$$

among them, z_l is a residual connection and the system would have a stable behavior for its stabilized gradient, i.e. non-vanishing or exploding during the time step of gradient transmission. When we densely use residual connections build a deeper neural network as follows

$$z_{l+1} = \mathbf{H}_l(z_0, z_1, ..., z_l) \tag{5}$$

where $\mathbf{H}_l$ is known as dense connections, the stabilization of the model is strengthened with better fitting performance. So that a deeper neural network with stabilized gradient propagating can be realized. From another point of view, raw features in shallow layers are still needed in deep layers as a reference or anchor for high-level information extraction, without those build in identity maps, the system would have to learn quasi-identity maps to pass those shallow information. It is not easy to learning and maintains such quasi-identity maps, which are full of uncertainties due to the randomly initialed neural weight and gradient noise.

Another important mechanism for DRC-NET to achieve excellent performance could be the inplace characteristics of the dense block that is processing feature maps without changing its size on frequency dimension. As our observation in the inplace model [15], the inplace characteristics could effectively preserve spatial cue, which delicately exist in local frequency bins. If that fine structure could be preserved, so do other details in T-F domains. For conveniently concatenate sequential features, there is no down-sampling convolution involved in the dense block, so the fine structure could be processed without down-sampling distortion. Besides, the down-sampling operation in DRC-NET does not increase channel size, which is different from the conventional approach which would increase channel size. It is because the local fine structure has been sufficiently processed by the dense block, so the down-sampling convolution only needed to extract a subset of feature map which embedded information with more sparsity, for it is getting more

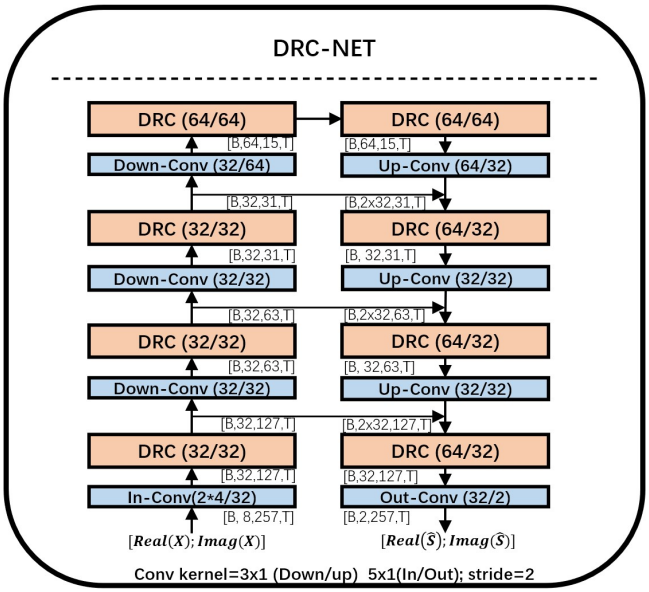

Fig. 2. The implemented DRC-NET framework

global and general due to the down-sampling operation.

2.4. DRC-NET constructions

The DRC-NET is based on the classical U-Net framework and is mainly constructed by the DRC blocks. Fig. 2 shows its architecture. For the DRC(i/o), i is the input channel, o is the output channel of the DRC block as well as the inside RC unit output channel. The Down-Conv(i/o) and Up-Conv(i/o) denote the convolution and transposed convolution with kernel size of 3x1, and stride of 2. Every convolutional layer in DRC-NET, except its input and output layer, is followed by ELU activation and layer normalization.

A drastic change from the DenseNET framework in [21] to the DRC-NET is that we remove the pyramid-like U-Net structure between the final dense block in the encoder and the first dense block in the decoder. Those structures are once very important in CRN model, by completely encoding frequency dimension into channel dimension, we could apply a single deterministic LSTM on it. But in DRC-NET, a massive amount of LSTM are already applied in channel-wise along the time dimension. And based on the previous works on inplace model [15], it is not required to completely encoding all frequency dimension in to channel dimension. So pruning those part is harmless, and because the channel size of those parts are very high, the pruning can significantly reduce the number of parameters as well as some computational complexity of the network.

3. EXPERIMENT AND EVALUATION

3.1. Experimental setup

In order to intuitively evaluate our models for dereverberation, we follow the experimental settings in [21], where both single-channel and 4-mic array configurations are considered.

Specifically, the WSJCAM0 [26] recording of the primary microphone is used as our speech corpus. The WSJCAM0 official dataset is divided into training, development and test parts. The utterances in the training set are clipped or duplicated to 4 seconds.

The room impulse responses (RIR) are simulated by image method [27]. The reverberation time (T60) is randomly selected

Table 1. The MAC and total parameters of different model

Model complexity		MAC(G)	Param(M)
Non-causal	DenseNET	14.4	17.8
	DenseCRN	15.9	18.0
	DRC-NET	13.9	2.6
Causal	DenseNET	4.7	6.2
	DenseCRN	5.3	6.3
	DRC-NET	7.5	1.4

from the range of $[0.2, 1.3]$s with uniform distribution, and the duration of RIR waveform is equal to its T60. The room length Rx and width Ry are randomly selected in the range of $[5, 10]$m, and room height Rz is in the range of $[3, 4]$m. A 4-channel uniform circular microphone array with 20 cm aperture is randomly placed at the room center, with random offsets of $[-0.5, 0.5]$m for each dimension. The first microphone angle ranges from $[0, \pi/4]$. The speaker is randomly placed on the same height plane with $[0.75, 2.5]$m away from array center. The early speech component of the first microphone is used as learning target. The average direct-to-reverberation energy ratio of the simulated reverberate speech corpus is -3.75 dB with a 4.7 dB standard deviation. The noise data set from REVERB challenge [28] is used as background noise. The noise is mixed with a reverberant speech by a randomly selected SNR at $[5, 25]$dB.

The sampling rate is set to 16 kHz, and we extract complex spectrum by using STFT with 32 ms window size of 8 ms frame shift. The real and imaginary parts of 257-dimension complex spectrum are concatenated as the model input. All models are trained by the Adam optimizer, with an initial learning rate of 0.001 and gradually tuning down by the verification loss. Mini-batch size is setting to 8.

The extended short-time objective intelligibility (ESTOI) [29], perceptual evaluation of speech quality (PESQ) [30], and the scale-invariant SDR (SI-SDR) [31] are used as evaluation metrics. Generally, the perception quality is depends on the clarity of the frequency fine structure, the speech intelligibility is relays on the time variation of frequency envelops, both of them are careless on phase distortion, and the SI-SDR gives consideration both for amplitude and phase in time domain.

3.2. Compared approaches

For evaluation and comparison, we implement following models.

i) DenseNET : the exact model proposed in [21].

ii) DenseNET(c) : the causal version of i), by setting the convolution kernel size in time dimension to 1, and using unidirectional LSTM.

iii) DenseCRN : our modified DenseNET, by replacing the middle convolutional layer in the dense block of i) DenseNET with a channel-wise LSTM on the time dimension, so it is dense CRN block inside, instead of dense CNN block.

iv) DenseCRN(c): the causal version of iii) DenseNET

v) DRC-NET : our implemented non-causal model.

vi) DRC-NET(c) : the causal version of DRC-NET, using unidirectional LSTM except in the RC(F) unit where using bidirectional LSTM modeling frequency patterns, and no convolution on the time dimension.

Their multiply-accumulate (MAC) and the number of parameter (Param) are shown in Table 1. To ensure the fairness of comparison, we keep the MAC or Param of the models at the same level.

4. EXPERIMENTAL RESULTS

Table 2 shows the comparison results of three methods with 4-mic and 1-mic configuration over the test set. It can be seen that the proposed two architectures, i.e. DenseCRN and DRC-NET, significantly and consistently outperform the SOTA DenseNET for different configurations (4mic/1mic and casual/non-casual). In particular, the performance of casual DRC-NET is even better than the non-casual DenseNET, with 90.3/89.8 for ESTOI, 3.40/3.37 for PESQ and 10.6/8.7 for SI-SDR.

Table 2. Comparisons between different approach and its causal version in the terms of ESTOI, PESQ, and SI-SDR

Metrics	ESTOI (%)		PESQ		SI-SDR	
Microphones	4mic	1mic	4mic	1mic	4mic	1mic
Unprocessed	48.7		1.95		-3.97	
DenseNET	89.8	84.7	3.37	3.12	8.7	6.5
DenseCRN	92.3	88.6	3.50	3.27	10.5	8.2
DRC-NET	**93.5**	**90.2**	**3.61**	**3.39**	**12.6**	**10.2**
DenseNET(c)	84.5	78.6	3.03	2.89	5.3	4.2
DenseCRN(c)	88.5	82.5	3.27	2.97	8.5	6.1
DRC-NET(c)	**90.3**	**85.3**	**3.40**	**3.09**	**10.6**	**7.4**

Specifically, comparing with DenseNET, the DenseCRN realized significant improvements on ESTOI, with an average 3.5(%) increased value in all 4 cases. This is because the channel-wise LSTM could significantly enhance the time sequence modeling capability, and the speech intelligibility is mainly reflected by the time variation of spectral. When it comes to the causal case, the monaural DenseCRN realize 0.08 improvement on PESQ, and a bigger 2.4 PESQ improvement is reported in 4-mic case, this indicates that the dense block with inplace nature and the channel-wise LSTM could effectively exploit spatial information, this result consistent with our previous conclusion in [15]. Besides, due to the absence of convolution on the time dimension, the causal DenseNET could barely benefit from a 4-microphone system with 0.13 PESQ promotion comparing with its monaural system, while 0.3 PESQ improvement can be achieved by our DenseCRN. This indicates that the channel-wise LSTM is a better choice for both spatial cue exploration and time sequence modeling. More importantly, with no damage to the causality and no frame-level latency requires, which is crucial for real-time application.

When we combine channel-wise LSTM and CNN into the RC unit, and alternately apply it along frequency and time dimension in the dense block, further impressive improvements are achieved by the DRC-NET. Specifically, in 4-channel configuration, the non-causal DRC-NET achieve 3.7(%) ESTOI, 0.24 PESQ, and 3.9 dB SI-SDR promotion comparing with DenseNET. In causal case, 5.8(%) ESTOI, 0.37 PESQ, and 5.3 dB SI-SDR promotion are realized by DRC-NET, and comparable promotions in the monaural case are also achieved. All those promotions are contributed by the RC unit, which combines finite and infinite neural response together, and realizes a better fitting performance. It should be noted that there are many utterances that already could be precisely recovered by DenseCRN with very high absolute metrics value, so relevant higher averaged metrics are more difficult for DRC-NET to reach. Besides, the parameter efficiency of DRC-NET is very impressive with only 2.6 million trainable parameters, which is very compact comparing with the DenseNET 17.8 million total parameters.

5. CONCLUSIONS

In this study, a massive amount of LSTM are distributed and combined with CNN in a densely connected neural network. Due to its combined infinite and finite neural response, the DRC-NET realized a powerful fitting performance in T-F domain speech dereverberation task. The experimental results confirm the effectiveness of those carefully designed mechanisms, which can be leveraged in further speech signal processing neural network design for various tasks.

6. REFERENCES

[1] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity-based blind dereverberation for single-channel speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 80–95, 2007.

[2] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[3] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[4] W. Liu and S. Weiss, *Wideband Beamforming — Concepts and Techniques*. 2010.

[5] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the mvdr beamformer in room acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 158–170, 2010.

[6] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, *et al.*, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge," in *Proceedings of REVERB Challenge Workshop, o2.3*, 2014.

[7] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. yu, W. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 436–443, IEEE, 2015.

[8] C. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Interspeech 2021 deep noise suppression challenge," in *INTERSPEECH 2021*, pp. 2796–2800, 2021.

[9] H. Li, X. Zhang, and G. Gao, "Robust speech dereverberation based on wpe and deep learning," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 52–56, IEEE, 2020.

[10] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1492–1501, 2017.

[11] Y. Liu, H. Zhang, X. Zhang, and L. Yang, "Supervised speech enhancement with real spectrum approximation," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5746–5750, 2019.

[12] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 9458–9465, 2020.

[13] Y. Luo and N. Mesgarani, "TasNet: Time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 696–700, IEEE, 2018.

[14] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 46–50, IEEE, 2020.

[15] J. Liu and X. Zhang, "Inplace Gated Convolutional Recurrent Neural Network for Dual-Channel Speech Enhancement," in *INTERSPEECH 2021*, pp. 1852–1856, 2021.

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

[17] K. Tan, X. Zhang, and D. Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5751–5755, IEEE, 2019.

[18] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2019.

[19] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, 2017.

[20] Z. Q. Wang and D. Wang, "Deep learning based target cancellation for speech dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 941–950, 2020.

[21] Z. Q. Wang and D. Wang, "Multi-microphone complex spectral mapping for speech dereverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 486–490, 2020.

[22] H.-S. Choi, S. Park, J. H. Lee, H. Heo, D. Jeon, and K. Lee, "Real-time denoising and dereverberation with tiny recurrent u-net," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5789–5793, IEEE, 2021.

[23] 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," *arXiv preprint arXiv:2008.00264*, 2020.

[24] Proakis, G. John, D. Manolakis, and G. Dimitris, *Digital Signal Processing: Principles, Algorithms and Applications*. 1992.

[25] W. Ee, "A proposal on machine learning via dynamical systems," *Communications in Mathematics and Statistics*, vol. 5, pp. 1–11, 2017.

[26] J. Fransen, D. Pye, T. Robinson, P. Woodland, and S. Young, "WSJCAM0 corpus and recording description," *Cambridge University Engineering Department (CUED), Speech Group, Trumpington Street, Cambridge CB2 1PZ, UK, Tech. Rep. CUED/F-INFENG/TR*, vol. 192, 1994.

[27] E. Habets, "Room impulse response generator," 2006.

[28] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the reverb challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *Journal on Advances in Signal Processing*, vol. 2016, 2016.

[29] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 1–1, 2016.

[30] 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 *IEEE International Conference on Acoustics*, 2002.

[31] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, 2019.