INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Rethinking Complex-Valued Deep Neural Networks for Monaural Speech Enhancement

[12]*Haibin Wu,* [1]*Ke Tan,* [1]*Buye Xu,* [1]*Anurag Kumar,* [1]*Daniel Wong*

[1]Meta Reality Labs Research, USA
[2]Graduate Institute of Communication Engineering, National Taiwan University
f07921092@ntu.edu.tw, {tanke1116, xub, anuragkr90, ddewong}@meta.com

Abstract

Despite efforts made to adopt complex-valued deep neural networks (CVDNNs), it remains unclear whether CVDNNs are generally more effective than real-valued DNNs (RVDNNs) for speech enhancement. This study systematically examines CVDNNs against their real-valued counterparts in monaural scenarios. We first investigate atomic units of CVDNNs against those of RVDNNs. We find the use of complex-valued operations hinders model capacity when model size is small. Moreover, we show that two notable CVDNNs, deep complex convolutional recurrent network (DCCRN) and deep complex U-Net (DCUNET), produce identical performance to their real-valued counterparts while requiring more computation. Our experimental results show that those CVDNNs do not provide a performance gain over RVDNNs for monaural speech enhancement, and are less desirable due to higher computational cost. This study suggests that it is more than nontrivial to rethink the efficacy of CVDNNs for speech enhancement.

Index Terms: Monaural speech enhancement, complex-valued neural networks, computational cost, deep learning

1. Introduction

Recent years have witnessed promising performance improvement of monaural speech enhancement models in the complex domain, given the importance of phase for speech quality [1–9]. A recent study [10] develops the key atomic components for complex-valued DNNs and claim that complex-valued parameters have various merits from computational, biological, and signal processing perspectives. Complex-valued DNNs, which operates with complex-valued arithmetic, seems to be advantageous for complex-domain speech enhancement, where DNNs are trained to learn complex spectrograms. Motivated by such an intuition, multiple efforts [3, 7, 11–16] adopted complex-valued DNNs for monaural speech enhancement. However, to the best of our knowledge, none of these studies has justified a performance gain provided by complex-valued DNNs over their real-valued counterparts with the same network structure and model size. Drude et al. [17] compared real- and complex-valued DNNs with fully-connected layers for beamforming, and found that the complex-valued DNN does not yield superior performance to the real-valued DNN while being more computationally expensive. Despite the promising performance improvement produced by recent complex-valued DNNs for monaural speech enhancement, it remains an open question whether it is the complex-valued nature that fundamentally brings the merits. Since the efficacy of complex-valued DNNs is likely different for single- and multi-microphone scenarios due to the availability of inter-channel phase relations, we focus on monaural speech enhancement in this investigation.

A recent notable model DCCRN [7] extends the convolutional recurrent network in [18] by replacing convolutional and long short-term memory (LSTM) layers with their complex-valued counterparts to estimate the ideal complex ratio mask. The DCCRN exhibits competitive performance over earlier works, which has drawn the community's attention to the efficacy of complex-valued DNNs for speech enhancement. However, we believe that it is premature to ascribe the performance improvement to the use of complex-valued operations due to the lack of systematic comparisons between DCCRN and its real-valued counterpart, in which only the complex-valued layers are replaced by the corresponding real-valued layers while all other configurations remain unaltered, including input features, training targets, training objectives, network structure and model size. Without such apples-to-apples comparisons, it is difficult to justify the attribution of the improvement achieved by complex-valued DNNs.

This study presents a critical assessment by systematically examining complex-valued DNNs against their real-valued counterparts through comprehensive comparisons:

1). Based on the principles of complex-valued computation [10], we formulate complex-valued DNN atomic units for investigation, including linear layers, convolutional layers, LSTM, and gated linear units. We compare their performance with that of their real-valued counterparts.

2). To comprehensively investigate complex-valued operations in different types of layer topology, we adopt gated convolutional recurrent network (GCRN) [5] - a real-valued DNN originally developed for complex-domain speech enhancement, which integrates a variety of layer types. We enumerate different versions of fundamental building blocks of GCRN, and show how different computing mechanisms in basic blocks affect the performance. We see that the models with complex-valued components do not outperform the real-valued counterparts. In addition, given the fact that many real-world applications require a computationally efficient model, we conduct the same comparisons with a setting where the model size is very small. We find that complex-valued operations even hinders enhancement performance relative to real-valued operations.

3). Two recent compelling models based on complex-valued operations, DCCRN [7] and DCUNET [3], have shown promising performance for monaural speech enhancement. In this work, we evaluate their real-valued versions with the same parameter count, and conduct investigation with different loss functions, learning rates and minibatch sizes, in terms of both enhancement performance and training stability. Evaluation results show that the complex-valued versions do not outperform their real-valued counterparts while they have higher computa-

This work was done while H. Wu was a research intern at Meta.

10.21437/Interspeech.2023-686

tional costs, which is consistent with the observation in [19].

2. Methodology

This section introduces the basic building blocks for complex-valued DNNs, followed by the case study design.

2.1. Building blocks

Linearity Fully connected layers, convolution layers and deconvolution layers are composed of matrix multiplications. We omit the bias to simplify the description. Taking the input complex-valued feature matrix as $X = X_r + jX_i$ and the complex-valued parameter matrix as $W = W_r + jW_i$, the complex-valued output can be elaborated as:

$$Y = (X_r W_r - X_i W_i) + j(X_r W_i + X_i W_r), \quad (1)$$

where Y denotes the output feature of the complex-valued layer, the subscripts r and i denote real and imaginary parts.

Activation function Given a complex-valued representation z, the activation function operates on the real and imaginary part independently as:

$$a = f(Re\ z) + jf(Im\ z), \quad (2)$$

where a is the output representation, Re and Im get the real and imaginary parts respectively, and f is the activation function.

LSTM For LSTM layers, we have two versions:

Quasi complex-valued LSTM In [7], the complex LSTM operation is treated as two separate operations on the real and imaginary parts. To be specific, they initialize two real-valued sub-LSTM layers, namely LSTM_r and LSTM_i, corresponding to the real and imaginary LSTM respectively. Given the input feature $X = X_r + jX_i$, the output feature can be derived as:

$$\begin{aligned} &F_{rr} = \text{LSTM}_r(X_r), F_{ir} = \text{LSTM}_r(X_i), \\ &F_{ri} = \text{LSTM}_i(X_r), F_{ii} = \text{LSTM}_i(X_i), \\ &F_{out} = (F_{rr} - F_{ii}) + j(F_{ri} + F_{ir}), \end{aligned} \quad (3)$$

where F_{out} is the output feature.

Fully complex-valued LSTM In addition to the quasi complex-valued LSTM, which does not perform complex-valued operations within sub-LSTM layers, we also investigate fully complex-valued LSTM, which totally follows the the arithmetic of complex numbers. Each matrix multiplication and activation function in this LSTM strictly follows the arithmetic in Sections 2.1 and 2.1.

Gated linear unit As adopted in [5], gated linear unit [20] is a widely used layer topology, consisting of two separate convolutional layers and one gating operation. The two separate convolutional layers process the same input, and produce their outputs $F^{(1)}$ and $F^{(2)}$, respectively. A sigmoid function is applied to $F^{(2)}$ to dervie a gate, which is then element-wisely multiplied with $F^{(1)}$ to yield the output of the gated linear unit. In a complex-valued gated linear unit, let $F^{(1)} = F_r^{(1)} + jF_i^{(1)}$ and $F^{(2)} = F_r^{(2)} + jF_i^{(2)}$ be the outputs of the two convolutional layers. We derive two gating mechanisms, i.e. separate gating and magnitude gating.

Separate gating We apply a sigmoid function to the real and imaginary parts of $F^{(2)}$ separately, which amounts to a complex-valued gate. The real and imaginary parts of this gate are element-wisely multiplied with $F_r^{(1)}$ and $F_i^{(1)}$, respectively.

Magnitude gating Unlike separate gating, magnitude gating calculates a real-valued gate $F^{(g)}$ from the magnitude of the complex feature map $F^{(2)}$:

$$F^{(g)} = (\sigma(|F^{(2)}|) - 0.5) \times 2, \quad (4)$$

where σ denotes the sigmoid function, and $|\cdot|$ extracts the magnitude of a complex feature map. Since the magnitude is non-negative, applying the sigmoid function to the magnitude always results in values ranging from 0.5 to 1. Hence we use an affine transformation to normalize the gating value to the range of 0 to 1. The resulting gate is applied to both real and imaginary parts of $F^{(1)}$, preserving the phase of $F^{(1)}$ [21].

2.2. Case study design

This section carefully designs the case studies, and elaborates the rationales and objectives of each case study. All pairs of real- and complex-valued models for comparison have the same configurations, including input features, training targets, training objectives, network structure and model size.

Basic Unit This case study compares different complex layers defined in Section 2.1 with their real-valued counterparts, in terms of enhancement performance and computational costs. Specifically, we compare: 1) a model with a stack of three complex-valued linear layers and its corresponding real-valued model, where each of the two hidden layers has 406 units in the complex-valued model and 512 units in the real-valued model, respectively. Such a configuration ensures that the two models have almost the same number of parameters. Note that each hidden layer is followed by a rectified linear unit function; 2) quasi complex-valued LSTM, fully complex-valued LSTM, and real-valued LSTM, each of which contains three LSTM layers followed by a linear output layer. In these three models, each LSTM layer contains 732, 732 and 1024 units, respectively. The implementations described in Section 2.1 are adopted for quasi complex-valued LSTM and fully complex-valued LSTM; 3) DCUNET, a convolutional encoder-decoder model developed in [3], and its real-valued counterpart (RUNET), in which all complex-valued convolutional, deconvolutional and linear layers are replaced by their real-valued counterparts. Akin to 1) and 2), we slightly adjust hyperparameters (e.g. number of out channels in convolutional layers) for RUNET, such that its model size is almost the same as DCUNET. Note that all these models are trained to learn complex spectral mapping.

GCRN GCRN [5] is a representative model for our investigation, because it consists of different types of layers including convolutional/deconvolutional layers, gated linear units, LSTM layers, and linear layers. The original GCRN has two decoders, one for real part estimation and the other for imaginary part estimation. We instead use a single shared decoder for both real and imaginary parts, corresponding to two output channels in the last deconvolutional layer of the decoder. Such an architecture can be naturally converted into complex-valued versions for comparison by replacing each layer with their complex-valued counterpart. In this case study, we aim to investigate: 1) whether replacing specific layers of GCRN with their complex-valued counterparts can lead to better performance; 2) how the use of complex-valued operations affect speech enhancement performance when the model is constrained to a relatively small amount of parameters; 3) which gating mechanism in Section 2.1 is the better choice, from both training stability and enhancement performance aspects. Note that regarding the bottleneck LSTM in GCRN, we adopt the quasi complex-valued LSTM for investigation.

Table 1: *Investigation of different basic units. The number of multiply-accumulate (MAC) operations is measured on a 1-second signal.*

	SNR	Noisy	(1a).C-LSTM	(1b).Quasi C-LSTM	(1c).LSTM	(1d).C-Linear	(1e).R-Linear	(1f).DCUNET	(1g).RUNET
STOI	-5 dB	0.69	0.85	0.86	0.86	0.61	0.61	0.85	0.85
	0 dB	0.78	0.90	0.91	0.91	0.70	0.70	0.90	0.90
	5 dB	0.85	0.94	0.94	0.94	0.76	0.76	0.94	0.94
WB-PESQ	-5 dB	1.11	1.65	1.71	1.69	1.12	1.12	1.64	1.70
	0 dB	1.15	1.95	2.02	2.00	1.17	1.28	1.92	2.00
	5 dB	1.24	2.29	2.35	2.34	1.24	1.25	2.27	2.36
SI-SDR (dB)	-5 dB	-5.00	10.80	11.10	10.87	0.92	1.23	10.80	10.87
	0 dB	0.05	13.62	13.92	13.78	4.69	4.94	13.79	13.86
	5 dB	5.01	16.36	16.64	16.55	7.19	7.57	16.74	16.84
# Para	-	-	23.35 M	23.35 M	23.62 M	0.59 M	0.59 M	3.10 M	3.12 M
# MACs	-	-	5.90 G	5.90 G	2.98 G	119.59 M	59.88 M	56.69 G	19.87 G

Table 2: *Investigation of different complex-valued components in GCRN. ♣, ♢, ♡ denote using the quasi complex-valued LSTM in the bottleneck, complex-valued convolutional layers, complex-valued deconvolutional layers, respectively. "- Separate" and "- Magnitude" denote using separate and magnitude gating mechanisms in GLUs, respectively, and ⊙ denotes the model performs complex ratio masking rather than complex spectral mapping originally used in [5].*

		STOI			**WB-PESQ**			**SI-SDR (dB)**			**# Para**	**# MACs**
		-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB		
	Noisy	0.69	0.78	0.85	1.11	1.15	1.24	-5	0.01	5.01	-	-
(2a)	GCRN (real-valued model)	0.84	0.90	0.94	1.57	1.87	2.24	8.30	11.29	14.13	9.25 M	1.72 G
(2b)	GCRN + ♣	0.83	0.90	0.94	1.55	1.85	2.22	8.22	11.17	13.98	9.25 M	2.57 G
(2c)	GCRN + ♢ - Separate	0.83	0.90	0.93	1.53	1.80	2.15	7.64	10.51	13.22	9.12 M	1.72 G
(2d)	GCRN + ♢ - Magnitude	0.83	0.90	0.94	1.56	1.85	2.23	7.66	10.63	13.43	9.12 M	1.72 G
(2e)	GCRN + ♣ + ♢ - Separate	0.83	0.90	0.93	1.52	1.81	2.16	8.14	11.16	14.02	9.12 M	2.57 G
(2f)	GCRN + ♣ + ♢ - Magnitude	0.84	0.90	0.94	1.56	1.87	2.24	7.89	10.89	13.76	9.12 M	2.57 G
(2g)	GCRN + ♢ + ♡ - Separate	0.83	0.89	0.93	1.53	1.83	2.20	7.95	10.96	13.88	8.83 M	1.72 G
(2h)	GCRN + ♢ + ♡ - Magnitude	0.83	0.90	0.94	1.54	1.85	2.23	7.67	10.75	13.79	8.83 M	1.72 G
(2i)	GCRN + ♣ + ♢ + ♡ - Separate	0.82	0.89	0.93	1.52	1.80	2.15	7.62	10.70	13.52	8.83 M	2.57 G
(2j)	GCRN + ♣ + ♢ + ♡ - Magnitude	0.83	0.90	0.94	1.57	1.88	2.27	7.65	10.87	13.82	8.83 M	2.57 G
(2A)	GCRN ⊙ (real-valued model)	0.83	0.89	0.93	1.50	1.79	2.16	7.28	10.33	13.40	9.25 M	1.72 G
(2J)	GCRN + ♣ + ♢ + ♡ - Magnitude ⊙	0.82	0.89	0.93	1.47	1.74	2.10	7.25	10.33	13.46	8.83 M	2.57 G

DCCRN In [7], the performance gain achieved by DCCRN is attributed by the authors to the complex multiplication constraint, which they believe can help DNNs learn complex representations more effectively. However, they did not compare DCCRN with its real-valued counterpart using the same configurations. Thus it is difficult to justify the attribution of the performance improvement, which is likely due to either the use of complex-valued operations or other components in the model design. The objective of this case study is to show whether DCCRN can outperform its real-valued counterparts, with the same amount of parameters. Specifically, we adopt the "DCCRN-E" configuration, which achieves the best performance in [7]. To derive the corresponding real-valued version, we simply replace the complex-valued layers with their real-valued counterparts, and reduce the channel numbers in the encoder to [32, 64, 64, 64, 128, 256] to maintain the number of parameters. Note that for consistency, we use the same training target (i.e. the ideal complex ratio mask) as in the original DCCRN paper [7].

3. Experiments

3.1. Experimental setup

The Interspeech2020 DNS Challenge training speech dataset [22] is used to create our training, validation and test sets, which contains roughly 65000 speech signals uttered by 1948 speakers in total. We randomly split these speakers into three distinct sets for training, validation and test sets, which include 1753 (∼90%), 97 (∼5%) and 98 (∼5%) speakers, respectively. Similarly, we partition the DNS Challenge noise dataset with around 65000 signals into 90%, 5% and 5% for training, validation and test sets, respectively. By randomly pairing speech and noise signals, we create a training set with 500000 noisy mixtures and a validation set with 1000 noisy mixtures, in both of which the signal-to-noise ratio (SNR) is randomly sampled between -5 and 5 dB. Following the same procedure, three test sets are created at different SNR levels, i.e. -5, 0 and 5 dB. Note that all speech and noise signals are randomly truncated to 10 seconds before mixing. We additionally use the synthetic test set released by DNS Challenge for evaluation.

All signals are sampled at 16 kHz. Short-time Fourier transform is performed to obtain spectrograms. We adopt the Adam optimizer to train all models. Multiple metrics are employed to measure the speech enhancement performance, including wideband perceptual evaluation speech quality (WB-PESQ) [23], short-time objective intelligibility (STOI) [24], scale-invariant signal-to-distortion ratio (SI-SDR) [25], DNSMOS P. 835 [26] and NORESQA-MOS [27].

Table 3: *Comparison between GCRN and CGCRN with relatively small model sizes. Subscripts "M" and "S" denote a middle size and a small size, respectively.*

	SNR	Noisy	CGCRN_M	GCRN_M	CGCRN_S	GCRN_S
STOI	-5 dB	0.69	0.81	0.81	0.75	0.79
	0 dB	0.78	0.88	0.88	0.83	0.86
	5 dB	0.85	0.92	0.92	0.88	0.91
WB-PESQ	-5 dB	1.11	1.39	1.42	1.29	1.35
	0 dB	1.15	1.62	1.64	1.47	1.54
	5 dB	1.24	1.93	1.95	1.71	1.82
SI-SDR (dB)	-5 dB	-5.00	6.60	7.25	4.14	5.82
	0 dB	0.05	9.75	10.26	6.80	8.94
	5 dB	5.01	12.58	12.94	8.73	11.85
# Para (M)	-	-	2.26	2.36	0.61	0.63
# MACs (M)	-	-	657.90	439.03	172.12	115.99

3.2. Experimental results

Basic Unit In Table 1, 1). columns (1a), (1b), (1c) denote the fully complex-valued LSTM, quasi complex-valued LSTM and real-valued LSTM. 2). Real-valued LSTM has half as many MACs as both complex-valued LSTMs. Among the three models, the quasi complex-valued LSTM achieves the best performance, while its improvement over the real-valued LSTM is marginal. Columns (1d) and (1e) denote the complex- and real-valued DNNs consisting of linear layers. Although the real-valued DNN only has half of the MAC number in the complex-

Table 4: *Comparisons between real- and complex-valued versions of DCCRN with different training objectives. "-Real" means the real-valued version of DCCRN. "-SISDR", "-L_1", "-MSE" denote using SI-SDR, L_1 and mean squared error (MSE) losses for training, respectively, where both L_1 and MSE losses are computed on the clean and estimated real, imaginary and magnitude spectrograms.*

	SNR	Noisy	DCCRN-SISDR	DCCRN-Real-SISDR	DCCRN-L_1	DCCRN-Real-L_1	DCCRN-MSE	DCCRN-Real-MSE
STOI	-5 dB	0.69	0.87	0.87	0.86	0.86	0.85	0.85
	0 dB	0.78	0.92	0.92	0.91	0.91	0.90	0.90
	5 dB	0.85	0.95	0.95	0.95	0.95	0.94	0.94
WB-PESQ	-5 dB	1.11	1.78	1.80	1.73	1.69	1.55	1.56
	0 dB	1.15	2.13	2.14	2.05	2.00	1.83	1.86
	5 dB	1.24	2.51	2.54	2.43	2.38	2.16	2.19
SI-SDR (dB)	-5 dB	-5.00	11.01	11.06	8.36	8.28	8.09	8.18
	0 dB	0.05	14.00	14.06	11.25	11.19	11.20	11.27
	5 dB	5.01	16.99	17.05	14.36	14.27	14.41	14.54

Table 5: *Comparisons between real- and complex-valued versions of DCCRN with different training objectives on the DNS Challenge synthetic test set without reverberation. The SNR of noisy speech ranges from 0 to 19 dB with an average of around 9.07 dB.*

	Noisy	DCCRN-SISDR	DCCRN-Real-SISDR	DCCRN-L_1	DCCRN-Real-L_1	DCCRN-MSE	DCCRN-Real-MSE
STOI	0.92	0.97	0.97	0.97	0.97	0.97	0.97
WB-PESQ	1.58	2.92	2.89	2.92	2.86	2.61	2.64
SI-SDR (dB)	9.23	19.60	19.54	17.11	17.13	17.33	17.55
DNSMOS (OVRL)	2.48	3.30	3.33	3.28	3.30	3.19	3.20
NORESQA-MOS	1.90	4.31	4.34	4.27	4.31	3.80	3.96
# Para	-	3.67 M	3.64 M	3.67 M	3.64 M	3.67 M	3.64 M
# MACs	-	14.38 G	4.84 G	14.38 G	4.84 G	14.38 G	4.84 G

valued DNN, it still produces slightly better performance than the latter. 3). (1f) and (1g) denote the DCUNET and its corresponding real-valued version respectively We see that the real-valued UNET outperforms DCUNET in terms of both enhancement performance and computational efficiency.

GCRN In Table 2, (2a) is the original real-valued GCRN. (2b)-(2j) are the models where some components are replaced by the corresponding complex-valued version. Moreover, (2A) and (2J) have the same model structure as (2a) and (2j), but are trained to perform complex ratio masking rather than complex spectral mapping. In Table 3, we reduce the model size to roughly 2 M and 0.6 M, where "CGCRN" denotes the same configuration as (2j). We can observe: 1). Replacing the components of GCRN with their complex-valued versions can not get any performance gain, as shown in (2a)-(2j). 2). In the comparison between the models trained for complex ratio masking, i.e. (2A) and (2J), the real-valued model performs slightly better than the complex-valued model. 3). Although the magnitude gating and separate gating lead to similar performance, the training loss curve of the former is much more stable than that of the latter. It is likely because the magnitude gating preserves phase information which could help stabilize the training. 4). In the small model setting, the real-valued models consistently outperforms the complex-valued counterparts, and performance gaps increase as the model size becomes smaller.

DCCRN Tables 4 and 5 compare the DCCRN with its real-valued counterpart on our simulated test set and the DNS Challenge synthetic test set, respectively. The following observations are obtained: 1). With three different training objectives, i.e. SI-SDR, L_1 and MSE, the real- and complex-valued models yield almost identical performance in all the metrics on both datasets. Take, for example the -5 dB case with the SI-SDR training loss in Table 4. The STOI, WB-PESQ and SI-SDR improvements over noisy mixtures are 0.18, 0.67 and 16.01 dB for the complex-valued model, and 0.18, 0.69 and 16.06 dB for the real-valued model, respectively. 2) As shown in Table 5, the real-valued model produces slightly better scores than the complex-valued model in both DNSMOS and NORESQA-MOS, i.e. two metrics that highly correlate with subjective quality scores. 3). We have also made comparisons under settings with different learning rates and minibatch sizes. We find that DCCRN is less robust than its real-valued counterpart against different learning rates. In addition, both models produce very similar performance with different minibatch sizes. However, we do not show these comparison results due to the page limit. 4). The real-valued model has only one-third of the MAC amount in the complex-valued model. Specifically, the number of MACs for the complex-valued model is 14.38 G, while it is only 4.84G for the real-valued model. Given that the two models yield almost the same performance, the complex-valued model is less efficient for real-world applications.

4. Concluding remarks

Through the extensive experiments, we have the following findings for monaural speech enhancement: 1). Complex-valued DNNs yield similar performance to their real-valued counterparts with the same number of parameters. 2). When the model size is relatively small, the use of complex-valued operations is detrimental to the enhancement performance. 3). The performance gain achieved by DCCRN and DCUNET is not attributed to the use of complex-valued operations. Furthermore, complex-valued DNNs require more MACs than their real-valued counterparts, without any performance gain.

A complex number multiplication can break into four real number multiplications. Based on our systematic comparisons, we believe that real-valued DNNs have the capacity to achieve comparable performance to their complex-valued counterparts with the same model size and network structure. Although complex-valued DNNs intuitively seem a more natural choice than real-valued DNNs for processing complex spectrograms, they are more computationally expensive and thus an inferior choice for real applications that are efficiency-sensitive. We believe that there is no sufficient evidence justifying the superiority of complex-valued DNNs over real-valued DNNs for monaural speech enhancement. This study demonstrates that it is more than nontrivial to rethink the efficacy of complex-valued operations in speech enhancement systems.

5. References

[1] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2015.

[2] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *IEEE 27th International Workshop on Machine Learning for Signal Processing*. IEEE, 2017, pp. 1–6.

[3] 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.

[4] K. Tan and D. L. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 6865–6869.

[5] ——, "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.

[6] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, "A fully convolutional neural network for complex spectrogram processing in speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 5756–5760.

[7] 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 *Interspeech*, 2020, pp. 2472–2476.

[8] R. Xu, R. Wu, Y. Ishiwaka, C. Vondrick, and C. Zheng, "Listening to sounds of silence for speech denoising," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9633–9648, 2020.

[9] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1829–1843, 2021.

[10] C. Trabelsi, O. Bilaniuk, Y. Zhang, S. S. Dmitriy Serdyuk, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," in *6th International Conference on Learning Representations*, 2018.

[11] S. Lv, Y. Hu, S. Zhang, and L. Xie, "DCCRN+: Channel-wise subband DCCRN with SNR estimation for speech enhancement," in *Interspeech*, 2021, pp. 2816–2820.

[12] S. Lv, Y. Fu, M. Xing, J. Sun, L. Xie, J. Huang, Y. Wang, and T. Yu, "S-DCCRN: Super wide band DCCRN with learnable complex feature for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 7767–7771.

[13] S. Zhao, T. H. Nguyen, and B. Ma, "Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2021, pp. 6648–6652.

[14] Y. Sun, L. Yang, H. Zhu, and J. Hao, "Funnel deep complex u-net for phase-aware speech enhancement." in *Interspeech*, 2021, pp. 161–165.

[15] K. N. Watcharasupat, T. N. T. Nguyen, W.-S. Gan, S. Zhao, and B. Ma, "End-to-end complex-valued multidilated convolutional neural network for joint acoustic echo cancellation and noise suppression," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 656–660.

[16] H. Zou and J. Zhu, "DCTCN: Deep complex temporal convolutional network for real time speech enhancement," in *11th International Conference on Intelligent Control and Information Processing*. IEEE, 2021, pp. 112–118.

[17] L. Drude, B. Raj, and R. Haeb-Umbach, "On the appropriateness of complex-valued neural networks for speech enhancement." in *Interspeech*, 2016, pp. 1745–1749.

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

[19] A. Pandey and D. L. Wang, "Exploring deep complex networks for complex spectrogram enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 6885–6889.

[20] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 933–941.

[21] D. Hayakawa, T. Masuko, and H. Fujimura, "Applying complex-valued neural networks to acoustic modeling for speech recognition," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2018, pp. 1725–1731.

[22] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Interspeech*, 2020, pp. 2492–2496.

[23] I. Rec, "P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union, CH–Geneva*, 2005.

[24] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

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

[26] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 886–890.

[27] P. Manocha and A. Kumar, "Speech quality assessment through MOS using non-matching references," in *Interspeech*, 2022, pp. 654–658.