End-to-End Speech Separation Using Orthogonal Representation in Complex and Real Time-Frequency Domain

Kai Wang[1], Hao Huang[1], Ying Hu[1], Zhihua Huang[1], Sheng Li[2]*

[1]School of Information Science and Engineering, Xinjiang University, Urumqi, China
[2]National Institute of Information and Communications Technology (NICT), Kyoto, Japan

`terry_wang@stu.xju.edu.cn, hwanghao@gmail.com`

Abstract

Traditional single channel speech separation in the time-frequency (T-F) domain often faces the problem of phase reconstruction. Due to the fact that the real-valued network is not suitable for dealing with complex-valued representation, the performance of the T-F domain speech separation method is often constrained from reaching the state-of-the-art. In this paper, we propose improved speech separation methods in both complex and real T-F domain using orthogonal representation. For the complex-valued case, we combine the deep complex network (DCN) and Conv-TasNet to design an end-to-end complex-valued model. Specifically, we incorporate short-time Fourier transform (STFT) and learnable complex layers to build a hybrid encoder-decoder structure, and use a DCN based separator. Then we present the importance of weights orthogonality in the T-F domain transformation and propose a multi-segment orthogonality (MSO) architecture for further improvements. For the real-valued case, we performed separation in real T-F domain by introducing the short-time DCT (STDCT) with orthogonal representation as well. Experimental results show that the proposed complex model outperforms the baseline Conv-TasNet with a comparable parameter size by 1.8 dB, and the STDCT-based real-valued T-F model by 1.2 dB, showing the advantages of speech separation in the T-F domain.

Index Terms: speech separation, single channel, deep complex network, DCT, end-to-end learning

1. Introduction

Single-channel speech separation is challenging because only limited information is provided, and is currently a topic of great interest. Recently, deep learning (DL) based speech separation methods have achieved promising results. In early DL-based studies, the magnitude spectrum of the mixture in the time-frequency (T-F) domain was separated [1–6], and then the mixture phase was used for reconstructing the waveforms of the estimated sources, however the corrupted phase limited performance. Some improved phase reconstruction algorithms [7, 8] with additional phase models or processes have been proposed. Researchers have performed speech separation with the time domain signal [9–11] directly, which uses a learnable encoder and decoder to replace the fixed T-F domain transformation. However, the specific space generated in time domain methods lacks interpretability and the performance is unstable in extreme conditions [12–14]. Another solution to the phase problem is to exploit the complex-valued representation from the short-time Fourier transform (STFT), which belongs to the T-F domain separation branch [15–18]. However, these approaches achieve

separation using real-valued models, failing to follow complex arithmetic rules, and might lose some implicit information.

One possible approach to simultaneously model both the magnitude and phase in T-F domain speech separation is to use the deep complex network (DCN) for dealing with complex-valued representation. Trabelsi et al. [19] proposed elementary building blocks for the DCN. Some researchers have applied the DCN to image reconstruction [20], image classification [21] and automatic music transcription [22] tasks, and have achieved promising performance. A recent new model called the deep complex convolution recurrent network (DCCRN) [23] applies the DCN to achieve state-of-the-art performance in the speech enhancement task. Another approach to T-F domain speech separation is to avoid the complex-valued representation, e.g, the short-time discrete cosine transform (STDCT) can transform the waveform to the real-valued T-F domain. The discrete cosine transform (DCT) [24] uses cosine function as the basis. In [25], a warped discrete cosine transform (WDCT)-based approach was proposed to enhance the degraded speech.

In this paper, we examine the advantages of speech separation in both complex and real T-F domain by exploiting the well-known Conv-TasNet [11] as the baseline. For the complex-valued case, we propose deep complex Conv-TasNet (DC-Conv-TasNet), which combines the DCN and Conv-TasNet. More specifically, we incorporate differentiable STFT/iSTFT and learnable complex layers to propose a hybrid encoder-decoder structure. For the separation module, we borrow the structure of Conv-TasNet and implement each layer based on complex operators. Differentiability allows backward propagation, which includes STFT/iSTFT in the end-to-end system. And we dissect the gains of DC-Conv-TasNet by gradually replacing components of Conv-TasNet. Then we empirically show the importance of weights orthogonality in the T-F domain transformation, and we propose a multi-segment orthogonality (MSO) structure that utilizes advantages of the compound loss and the orthogonality of multi hybrid decoders to help training the segmented model. The proposed complex model outperforms the Conv-TasNet by 1.8 dB on SI-SNRi. For the real T-F domain, we introduce the differentiable STDCT in Conv-TasNet (named as DCT-Conv-TasNet). We also design the real hybrid encoder-decoder and multi-segment orthogonality (MSO) structure. The proposed real model outperforms the Conv-TasNet by 1.2 dB. The contributions of this work are summarized as follows:

• To the best of our knowledge, this is the first successful attempt in which the DCN is applied to the speech separation task in the complex T-F domain, and we also extend the method to the real T-F domain by introducing STDCT. We propose a hybrid encoder-decoder structure in the complex and real T-F domain respectively.

• We show that weights orthogonality in the T-F domain
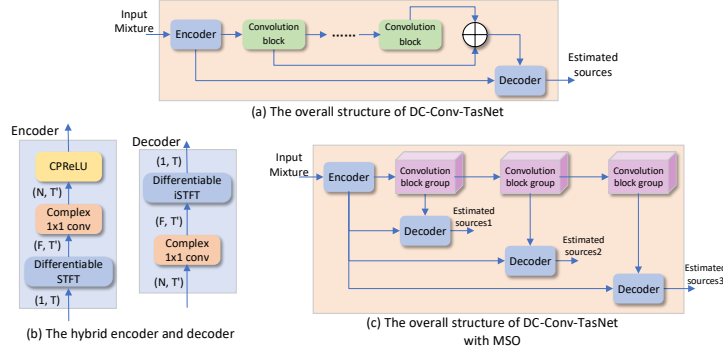
---

* Corresponding author.

Figure 1: *Structure of DC-Conv-TasNet. The convolution layers and activations in the overall structures are simplified.*

transformation is important to the improvements. We further propose multi-segment orthogonality (MSO) structure for further improvements. Experimental results show that the proposed complex and real models outperforms the baseline Conv-TasNet significantly, showing the superiority of speech separation in the T-F domain over the time domain.

2. DC-Conv-TasNet

2.1. Overall structure

Fig.1 (a) shows the overall structure of DC-Conv-TasNet, in which we design the hybrid encoder-decoder, and build the separator by replacing components of Conv-TasNet with corresponding complex-valued blocks. The original separator of Conv-TasNet is based on temporal convolutional network(TCN). For details of Conv-TasNet, please refer to [11]. The network configurations are listed in Table 1. Compared with Conv-TasNet, the hyper-parameters B and Sc are reduced from 128 to 96, and N and H from 512 to 384, which makes the model sizes comparable.

2.2. Hybrid encoder-decoder

The T-F domain transformation has the characteristics of weights orthogonality and signal reconstruction. On one hand, weights orthogonality brings gains for deep learning. The works [26, 27] developed novel orthogonality regularizations on training deep models. Povey et al. [28] improved speech recognition tasks by factorizing the parameter matrix into two smaller matrices with one constrained to be orthogonal. On the other hand, signal reconstruction is beneficial to speech separation that needs to reconstruct the estimated sources. Some two-stage methods [29, 30] for speech separation presented that pre-training an encoder-decoder structure which reconstructs the in-

put signal will improve the separation results. This pre-trained encoder could output sparse representation, which is similar to the T-F domain transformation.

However, the fixed parameters of the traditional T-F domain transformation are not learnable, that limit its potential. Heitkaemper et al. [14] showed that only using the STFT/iSTFT as the encoder/decoder would lead to poor separation results, and so it does in our experimental results. Instead of only using the STFT/iSTFT as the encoder/decoder, we design a hybrid encoder-decoder structure, as shown in Fig.1 (b). We add a complex 1×1 convolution layer and CPReLU activation [31] after STFT, and a complex 1×1 transposed convolution layer before iSTFT. The learnable layers can modulate the T-F domain representation to a specific space for the separator, while keeping the influence of orthogonality.

2.3. Multi-segment orthogonality (MSO)

Nachmani et al. [32] grouped the RNN-based model and evaluated the error after each RNN group, and showed that the compound loss was beneficial in the time domain. Based on the grouping and compound loss, we propose multi-segment orthogonality (MSO) in the T-F domain. The long distance between the learnable layers and the orthogonal transformations might weaken the the gains from orthogonality. We divide the original separator of DC-Conv-TasNet into groups and employ the hybrid decoder for each group, to improve the benefits from orthogonality.

The structure of DC-Conv-TasNet with MSO is shown in Fig.1 (c). We divide all the 24 stacked complex convolution blocks into three groups. Because the increasing dilation factors in the convolution blocks are repeated three times [11], we regard one repeat as one group. And then we add the complex 1×1 convolution, complex activation and hybrid decoder for each group to obtain separated sources.

2.4. Implementation of the complex-valued layers

We follow the work in [19] for the implementation of the complex-valued layers. We represent a complex vector with real-valued vectors. Consider a typical real-valued tensor that has $2N$ channels, to represent it as complex-valued, we assign the first N channels to represent the real components and the remaining N channels to represent the imaginary components.

Convolution layer: We represent a complex convolution layer using two real-valued convolution layers, one as the real part and the other as the imaginary part. Given a complex con-

Table 1: *Hyper-parameters of DC-Conv-TasNet*

Symbol	Value	Description
F	*varying*	Number of frequency channels of STFT
N	384	Number of output channels of the encoder
W	*varying*	Length of filters in STFT
Hop	8	Hop length of filters in STFT
B	96	Number of channels in the bottleneck
Sc	96	Number of channels in the skip-connection
H	384	Number of channels in convolution blocks
P	3	Kernel size in convolution blocks
X	8	Number of convolution blocks in each repeat
R	3	Number of repeats

volution filter $W = W_r + jW_i$ and the input vector $x = a + jb$, we can obtain the output:

$$W * x = (W_r * a - W_i * b) + j(W_r * b + W_i * a), \quad (1)$$

where $*$ denotes the convolution operator.

Activation: Several complex activations have been proposed in early literatures, e.g. MODReLU [33], CReLU [19], zReLU [34] and CPReLU [31]. We empirically found that CPReLU yielded a better performance for our model. Compared with CReLU, CPReLU can retain the information of representation with negative values.

Normalization: We apply complex layer normalization (CLN) for our model, by combining the layer normalization in Conv-TasNet [11] and the complex batch normalization [19]. For implementation details of CLN, please refer to [19].

2.5. Complex ratio mask

We seek to estimate the complex ratio mask (CRM) [15, 35] for each source, which is phase-aware and guaranteed to achieve promising performance. Assuming the estimated CRM of the cth source is $\tilde{M}_c = \tilde{M}_{c,r} + j\tilde{M}_{c,i}$, and the complex representation of the mixture is $X = X_r + jX_i$, we can obtain the the corresponding estimated source in complex field:

$$\tilde{S}_c = (\tilde{M}_{c,r}X_r - \tilde{M}_{c,i}X_i) + j(\tilde{M}_{c,i}X_r + \tilde{M}_{c,r}X_i) \quad (2)$$

Then we use the hybrid decoder to obtain the estimated source $\tilde{s}_c$ in the time domain.

3. DCT-Conv-TasNet

DCT-Conv-TasNet is a T-F domain separation model using real-valued network, by incorporating STDCT into Conv-TasNet. We stack an STDCT, 1×1 convolution layer and ReLU activation to build a real-valued hybrid encoder. For the hybrid decoder, we use an 1×1 transposed convolution layer and inverse STDCT (iSTDCT), i.e.: $x_{en} = \mathrm{ReLU}(\mathrm{Conv}(\mathrm{STDCT}(x)))$ and $x_{de} = \mathrm{iSTDCT}(\mathrm{TransConv}(x))$. The separation module is inherited from Conv-TasNet. We also design MSO version of the model, DCT-Conv-TasNet-MSO, by segmenting the stacked convolution blocks of Conv-TasNet and adding corresponding components for each group, similar to that of DC-Conv-TasNet in Fig.1 (c).

4. Experiments and results

4.1. Experimental setup

We experiment with the proposed models on the WSJ0-2mix [1] dataset. 8 kHz sampling is used. The mixtures are generated by mixing two random utterances from different speakers in the Wall Street Journal dataset (WSJ0) with a random SNR between -5 dB and 5 dB. About 30 hours of training and 10 hours of validation speech data are generated from recordings in the training set si_tr_s from the WSJ0 dataset. We mix two random speakers from the WSJ0 development set si_dt_05 and evaluation set si_et_05 in the same manner to generate 5 hours of audio for the evaluation set.

We use negative scale-invariant source-to-noise ratio (SI-SNR) [11] as the objective function. For the model with MSO, we apply the loss:

$$\mathcal{L}(s, \{\hat{s}_m\}_{m=1}^M) = -\frac{1}{M}\sum_{m=1}^M \mathrm{SI\text{-}SNR}(\boldsymbol{s}, \hat{\boldsymbol{s}_m}), \quad (3)$$

where s represents the clean sources and $\{\hat{s}_m\}_{m=1}^M$ denote the estimated sources from the M groups. While in inference, considering that the later blocks will benefit from the results of the previous blocks, we only use the last group output $\hat{s}_M$.

For the STFT and STDCT, various window lengths are evaluated, and the hop length is fixed to 8 samples. The SI-SNR improvement (SI-SNRi) is used as the evaluation metric. We also report the perceptual evaluation of speech quality (PESQ [36]). Adam [37] is used as the optimizer, and the initial learning rate is set to 0.001. All the models are trained for 100 epochs on 4-second long segments. Gradient clipping with a maximum L2-norm of 5 is applied during training. Both the real and complex initial slopes in CPReLU are set to 0.25. We found empirically that the parameter initialization for DCN in [19] yields no improvement in our experiments, thus we initialize the parameter with the random settings.

4.2. Results of DC-Conv-TasNet

Table 2 presents the results of DC-Conv-TasNet. We investigated four window length settings: 16, 32, 64, 128. The larger number represents more fine-grained features. We reproduced Conv-TasNet [11] as the baseline. From Table 2, it can be observed that the proposed model with various window lengths outperforms the baseline in SI-SNRi and PESQ. The window length of 64 shows the best performance. The reason may be that too much channel information exceeds the representation capacity of the convolution layers in the model. The window length of 64 is fixed for the latter experiments of DC-Conv-TasNet. Table 2 also lists the oracle results, including the ideal binary mask (IBM), ideal ratio mask (IRM), ideal phase sensitive mask (IPSM), and complex ideal ratio mask (cIRM) [35]. As expected, the results of cIRM are almost perfect reconstruction, that provide a promising upper-bound for our method.

4.3. Dissection results of DC-Conv-TasNet

We dissect the DC-Conv-TasNet to demonstrate the contributions of each components. The results are shown in Table 3. Starting from Conv-TasNet, we first replace the learnable encoder/decoder with the STFT/iSTFT respectively (STFT+TCN), in which the real and imaginary parts of the complex input are concatenated along the frequency dimension to be processed in real-valued domain. We can see that the combination of STFT+TCN yields inferior results. Then we investigate the combination of the hybrid encoder-decoder and the TCN (Hybrid+TCN). The SI-SNRi is increased to 15.3 dB and PESQ to 3.35, showing the importance of the hybrid encoder-decoder, but still lower than the baseline. We can see the disadvantages of the real-valued network when it is operated on

Table 2: *Results of DC-Conv-TasNet*

Model	Window length	Model size	SI-SNRi (dB)	PESQ
DC-Conv-TasNet	16	5.8M	16.5	3.53
	32	5.8M	16.7	3.57
	64	5.8M	**16.8**	**3.59**
	128	5.8M	16.6	3.57
Conv-TasNet [1]	-	5.1M	15.7	3.40
IBM	-	-	13.3	3.49
IRM	-	-	12.2	3.93
IPSM	-	-	16.0	4.23
cIRM	-	-	63.3	4.50

[1] The result reported here is from our implementation.

Table 3: *Results from Conv-TasNet to DC-Conv-TasNet*

Model	STFT	Conv	CPReLU	NET	SI-SNRi(dB)	PESQ
Conv-TasNet					15.7	3.40
STFT+TCN	✓			TCN	14.0	3.19
Hybrid+TCN	✓	✓	✓	TCN	15.3	3.35
STFT+DCN	✓			DCN	15.7	3.49
STFT+CPReLU+DCN	✓		✓	DCN	15.8	3.47
STFT+Conv+DCN	✓	✓		DCN	16.6	3.56
DC-Conv-TasNet	✓	✓	✓	DCN	16.8	3.59

Table 4: *Results on orthogonality in T-F domain*

Model	Domain	Ortho.	MSO	SI-SNRi(dB)	PESQ
Conv-TasNet	Time			15.7	3.40
DC-Conv-TasNet-NO				16.3	3.53
DC-Conv-TasNet	Complex T-F	✓		16.8	3.59
DC-Conv-TasNet-MSO		✓	✓	17.5	3.65
DCT-Conv-TasNet-NO				15.7	3.39
DCT-Conv-TasNet	Real T-F	✓		16.0	3.44
DCT-Conv-TasNet-MSO		✓	✓	16.9	3.52

complex-valued inputs. Next we keep STFT and replace the TCN with the DCN, shown as STFT+DCN. The SI-SNRi is increased to 15.7 dB and PESQ to 3.49, showing that DCN is effective for the complex-valued representation. Finally we test by adding the complex 1×1 convolution layer and CPReLU respectively, shown as STFT+Conv+DCN (corresponding decoder is Conv+iSTFT) and STFT+CPReLU+DCN (corresponding decoder is iSTFT). When we add CPReLU, there is almost no improvement for the results, while STFT+Conv+DCN yields significant improvements, i.e. 16.6 dB in SI-SNRi and 3.56 in PESQ. Therefore, compared with CPReLU, complex 1×1 convolution layer plays a more important role in the hybrid encoder.

4.4. Results on orthogonality of DC-Conv-TasNet

Here we examine the contribution of orthogonality in DC-Conv-TasNet, as shown in the complex T-F domain related results in Table 4. We first remove the orthogonality by using the learnable complex 1-D convolution layers to replace the STFT/iSTFT respectively, denoted as DC-Conv-TasNet-NO (DC-Conv-TasNet with non-orthogonality). The SI-SNRi is decreased from 16.8 dB to 16.3 dB and PESQ from 3.59 to 3.53. Then we apply the MSO to DC-Conv-TasNet to enhance the influence of orthogonality. The SI-SNRi is increased to 17.5 dB and PESQ to 3.65, further showing the advantage of orthogonality for the spectral representation in the complex T-F domain.

4.5. Results of DCT-Conv-TasNet

We extend the above corresponding experiments in DC-Conv-TasNet to real-valued domain, i.e., DCT-Conv-TasNet. The results are shown in the lower part of Table 4. The window length is set to 32 to show the best results. We can see that the DCT-Conv-TasNet outperforms the baseline by 0.3 dB in SI-SNRi. Considering only a pair of STDCT/iSTDCT are added, it is still a potentially promising result. We report results of applying the MSO and removing the orthogonality by replacing the STDCT/iSTDCT with corresponding 1-D convolution layers in DCT-Conv-TasNet, denoted as DCT-Conv-TasNet-MSO and DCT-Conv-TasNet-NO, respectively. Removing orthogonality decreases the SI-SNRi and PESQ, while applying the MSO yields further improvements significantly. Similar to the
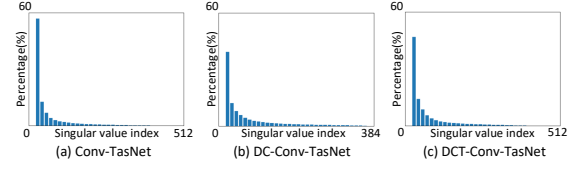

Figure 2: *Distribution of PCA singular values of the representation from encoders. The vertical axis represents the percentage of the singular value to the sum of all singular values.*

results in complex T-F domain, we can observe the advantage of real T-F domain speech separation, and weights orthogonality is an important factor.

4.6. Correlation of the representation

We carry out further analysis on the influence of weights orthogonality in the T-F domain transformation on the correlation of the representation. We perform principal component analysis (PCA) on the outputs of encoders in Conv-TasNet, DC-Conv-TasNet and DCT-Conv-TasNet, respectively. For DC-Conv-TasNet, we use the modulus of the complex representation for PCA. Fig.2 show the comparisons of the correlation via the distribution of singular values. It can be seen that the distribution in Conv-TasNet is more concentrated, and many singular values of greater indices are almost zero, showing high correlation between the representation vectors. While the concentration of distribution in DC-Conv-TasNet and DCT-Conv-TasNet is relatively lower. The speech separation task, that reconstructs the estimated signal accurately, requires more detailed information. The representation of T-F domain methods has low correlation, leading to a better capacity for modeling the detailed information.

5. Conclusions

In this paper, we have shown the advantages of speech separation in the complex and real T-F domain over the time domain. For the complex T-F case, we have applied the STFT-based hybrid encoder-decoder and Deep Complex Network (DCN) to the Conv-TasNet, and propose a novel multi-segment orthogonality (MSO) structure. The proposed complex model yields significant improvements. For the real T-F case, we have introduced STDCT in the Conv-TasNet, which is simple but also effective. We have empirically shown the importance of weights orthogonality in the T-F domain transformation for separation, that reduces the correlation of representation to be beneficial for reconstructing the estimated sources accurately. In addition, the proposed complex and real hybrid encoder-decoder and MSO structure can be generalized to other complex-valued and real-valued models in the T-F domain. Further work includes investigating T-F domain versions of other speech separation models that are more powerful, e.g. the dual-path RNN [38] and transformer-based methods [39, 40].

6. Acknowledgements

7. References

[1] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[2] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*, 2016, pp. 545–549.

[3] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*, 2017, pp. 241–245.

[4] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[5] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. ICASSP*, 2017, pp. 246–250.

[6] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.

[7] Z.-Q. Wang, K. Tan, and D. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," in *Proc. ICASSP*, 2019, pp. 71–75.

[8] Z. Ni and M. I. Mandel, "Mask-dependent phase estimation for monaural speaker separation," in *Proc. ICASSP*, 2020, pp. 7269–7273.

[9] S. Venkataramani, J. Casebeer, and P. Smaragdis, "End-to-end source separation with adaptive front-ends," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, 2018, pp. 684–688.

[10] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *Proc. ICASSP*, 2018, pp. 696–700.

[11] ——, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[12] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," in *Proc. Interspeech*, 2019, pp. 1368–1372.

[13] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "Whamr!: Noisy and reverberant single-channel speech separation," in *Proc. ICASSP*, 2020, pp. 696–700.

[14] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying tasnet: A dissecting approach," in *Proc. ICASSP*, 2020, pp. 6359–6363.

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

[16] F. Bahmaninezhad, J. Wu, R. Gu, S.-X. Zhang, Y. Xu, M. Yu, and D. Yu, "A comprehensive study of speech separation: spectrogram vs waveform separation," in *Proc. Interspeech*, 2019, pp. 4574–4578.

[17] Y. Liu and D. Wang, "Divide and conquer: A deep casa approach to talker-independent monaural speaker separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2092–2102, 2019.

[18] M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Filterbank design for end-to-end speech separation," in *Proc. ICASSP*, 2020, pp. 6364–6368.

[19] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. Pal, "Deep complex networks," *arXiv preprint arXiv:1705.09792*, 05 2017.

[20] E. Cole, J. Cheng, J. Pauly, and S. Vasanawala, "Analysis of deep complex-valued convolutional neural networks for mri reconstruction," *arXiv:2004.01738v4*, 2020.

[21] Y. Cao, Y. Wu, P. Zhang, W. Liang, and M. Li, "Pixel-wise polsar image classification via a novel complex-valued deep fully convolutional network," *Remote Sensing*, vol. 11, no. 22, p. 2653, 2019.

[22] M. Yang, M. Q. Ma, D. Li, Y.-H. H. Tsai, and R. Salakhutdinov, "Complex transformer: A framework for modeling complex-valued sequence," in *Proc. ICASSP*, 2020, pp. 4232–4236.

[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] N. N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. C-23, no. 1, pp. 90–93, 2006.

[25] J.-H. Chang, "Warped discrete cosine transform-based noisy speech enhancement," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 52, no. 9, pp. 535–539, 2005.

[26] N. Bansal, X. Chen, and Z. Wang, "Can we gain more from orthogonality regularizations in training deep cnns?" in *Proceedings of NIPS*, 2018, pp. 4266–4276.

[27] B. Liu, Y. Zhu, Z. Fu, G. de Melo, and A. Elgammal, "Oogan: Disentangling gan with one-hot sampling and orthogonal regularization," in *Proceedings of the AAAI*, vol. 34, no. 04, 2020, pp. 4836–4843.

[28] D. Povey, G. Cheng, Y. Wang, K. Li, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Interspeech 2018*, 2018.

[29] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan, and P. Smaragdis, "Two-step sound source separation: Training on learned latent targets," in *Proc. ICASSP*, 2020, pp. 31–35.

[30] H. Wang, Y. Song, Z.-X. Li, I. McLoughlin, and L.-R. Dai, "An online speaker-aware speech separation approach based on time-domain representation," in *Proc. ICASSP*, 2020, pp. 6379–6383.

[31] A. Pandey and D. Wang, "Exploring deep complex networks for complex spectrogram enhancement," in *Proc. ICASSP*, 2019, pp. 6885–6889.

[32] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7164–7175.

[33] M. Arjovsky, A. Shah, and Y. Bengio, "Unitary evolution recurrent neural networks," in *International Conference on Machine Learning*, 2016, pp. 1120–1128.

[34] N. Guberman, "On complex valued convolutional neural networks," *arXiv preprint arXiv:1602.09046*, 2016.

[35] Z. Wang, X. Wang, X. Li, Q. Fu, and Y. Yan, "Oracle performance investigation of the ideal masks," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016, pp. 1–5.

[36] 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*, 2001, pp. 749–752.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[38] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. ICASSP*, 2020, pp. 46–50.

[39] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.

[40] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," *arXiv preprint arXiv:2010.13154*, 2020.