FULLY COMPLEX DEEP NEURAL NETWORK FOR PHASE-INCORPORATING MONAURAL SOURCE SEPARATION

Yuan-Shan Lee[1], *Chien-Yao Wang*[1], *Shu-Fan Wang*[1], *Jia-Ching Wang*[1], *and Chung-Hsien Wu*[2]

[1]Dept. of Computer Science and Information Engineering, National Central University, Taiwan
[2]Dept. of Computer Science and Information Engineering, National Cheng Kung University, Taiwan

ABSTRACT

Deep neural network (DNN) have become a popular means of separating a target source from a mixed signal. Most of DNN-based methods modify only the magnitude spectrum of the mixture. The phase spectrum is left unchanged, which is inherent in the short-time Fourier transform (STFT) coefficients of the input signal. However, recent studies have revealed that incorporating phase information can improve the quality of separated sources. To estimate simultaneously the magnitude and the phase of STFT coefficients, this work paper developed a fully complex-valued deep neural network (FCDNN) that learns the nonlinear mapping from complex-valued STFT coefficients of a mixture to sources. In addition, to reinforce the sparsity of the estimated spectra, a sparse penalty term is incorporated into the objective function of the FCDNN. Finally, the proposed method is applied to singing source separation. Experimental results indicate that the proposed method outperforms the state-of-the-art DNN-based methods.

Index Terms— Deep neural network, phase information

1. INTRODUCTION

The human auditory system can segregate the interested source form the mxiture. For example, a person can easily focus on particular instruments when listening to classical music. Some investigations have demonstrated that the redundancy reduction is essential to mammalian perceptual processing [1, 2]. This work concerns source separation for the monaural music signal. Supervised approaches, which involve prior training [3–7], have recently been intensively investigated. Huang and Kim introduced a joint optimized DNN model for monaural source separation [6]. After the short-time Fourier transform (STFT) coefficients of signals were extracted, a DNN was used to learn the nonlinear mapping between input mixture and target sources. In general, DNN performed very well for musical signals under unseen noisy conditions [7].

The DNN-based methods [5–10] typically focus on modifying only the magnitude response of complex-vauled STFT cofficients, leaving the phase response unchanged in the separation process. However, recent research [11, 12] has established that the perceptual quality of separated sources can be improved by enhancing the phase spectrum. This fact leads some approaches to consider phase information in source separation [13–15]. Williamson *et al*. [14, 15] developed a DNN-based speech separation method which divide the complex-valued STFT coefficients into real and imaginary components. The two real-valued components were combined into one feature vector. In [16], the magnitude and phase components are also stacked into a signal vector. A standard DNN was then employed to learn the connection between the mixture and the sources in the real domain. Notably, the inherent structure of complex-valued STFT coefficients is changed. The number of neurons in the input and output layers are doubled. Accordingly, directly learning the model in the complex domain may be more natural.

In this paper, the source separation problem is addressed using a fully complex-valued DNN (FCDNN). The proposed method makes two important contributions. First, the developed FCDNN directly learns the nonlinear relationship between input music and target sources in a fully complex domain. Accordingly, the inherent structure of complex-valued data is maintained during the learning process. Second, a sparse objective for the proposed FCDNN, which enhances the sparsity of the estimated spectra, is investigated. The rest of this paper is arranged as follows. Section 2 review the previous works on DNN-based source separation. Section 3 presents the proposed FCDNN-based source separation in detail. Experimental setting and results are described in Section 4. Section 5 draws conclusions.

2. BACKGROUND

Most DNN-based approaches employ a data-driven scheme to solve the source separation problem. The DNN is adopted to supervisely learn the non-linear relationship between the input mixture and the target mask in the time-frequency domain. Given a discrete-time signal $x(t) \in \mathbb{R}$ that represents a mixture of P sources, $x(t) = \sum_{p=1}^{P} s_p(t)$, $t \in \mathbb{Z}^+$, the discrete Fourier transform (DFT) is applied to obtain the

This work was supported by the Ministry of Science and Technology, Taiwan, R.O.C., under Grant No. MOST 104-2628-E-008-002-MY3.

complex-valued spectrogram of $x(t)$ as

$$\mathbf{X} = \mathrm{STFT}\left(\mathbf{x}\right) = \{X(n,k)\}_{n,k\in\Omega} \in \mathbb{C}^{K\times N}, n,k \in \mathbb{N} \tag{1}$$

where $X(n,k)$ denotes the complex-valued STFT coefficients for every frequency bin k and time frame n. The term Ω represents the domain of (n,k) for $1 \leqslant n < N, 1 \leqslant k < K$. For simplicity, the subscripts n and k are subsequently omitted. The STFT coefficients can be further represented as product of two terms as follows,

$$X = |X|\, e^{i\phi_X} \in \mathbb{C} \tag{2}$$

where $|X| \in \mathbb{R}$ represents the magnitude of X, $i = \sqrt{-1}$ and $\phi_X \in \mathbb{R}$ is the phase information in a time-frequency bin. Similarly, the STFT coefficient of the p-th source can be represented by $S_p = |S_p|\, e^{i\phi_{S_p}}$. In the time-frequency domain, the estimated STFT coefficient of source S_p is denoted as $\hat{S}_p$. Nowadays, most DNN-based source separation methods [5–10] focus only on modification of the magnitude term as follows.

$$\hat{S}_p = \bar{S}_p e^{i\phi_X} \tag{3}$$

Herein, ϕ_X equals the phase term of the input mixture; $\bar{S}_p \in \mathbb{R}$ denotes the estimated magnitude of S_p, which can be extracted using the time-frequency mask M_p.

In DNN-based source separation, the ideal ratio mask (IRM) [5–7] is commonly used to separate the sources, which can be defined as,

$$M_p = \frac{|S_p|}{\sum_{p=1}^{P}|S_p|} \in \mathbb{R} \tag{4}$$

The STFT cofficient of the p-th source can be obtained using $\hat{S}_p = M_p X$. Finally, the separated source is obtained by applying the inverse short-time Fourier transform (iSTFT) to $\hat{S}_p$. However, a recent study [12] has report that the resynthesized spectrogram that is obtained using this manner is inconsistent, meaning that $\hat{S}_p \neq \mathrm{STFT}(\mathrm{iSTFT}(\hat{S}_p))$. Moreover, the phase estimation of source is not conducted, degrading the perceptual quality of separated sources.

To incorporate phase estimation into DNN-based source separation, Williamson *et al*. [14, 15] developed the complex-valued ideal ratio mask (C-IRM), which is defined as,

$$M_p^{\mathbb{C}} = \frac{S_p}{\sum_{p=1}^{P} S_p} = \frac{X^{\Re}S_p^{\Re} + X^{\Im}S_p^{\Im}}{(X^{\Re})^2 + (X^{\Im})^2} + i\cdot\frac{X^{\Re}S_p^{\Im} - X^{\Re}S_p^{\Im}}{(X^{\Re})^2 + (X^{\Im})^2} \tag{5}$$

where $(\cdot)^{\Re}$ and $(\cdot)^{\Im}$ are operators that extract real and complex components, respectively. The C-IRM provides additional phase information compared to the IRM. In [14] and [15], the complex-valued STFT coefficients and C-IRM were firstly divided into real and imaginary components. Then, the real-valued components were fed into the DNN as the input and target. Notably, the inherent structure of the complex-valued features was changed. The DNN weights were real-valued and could not represent spectral patterns in the complex domain.

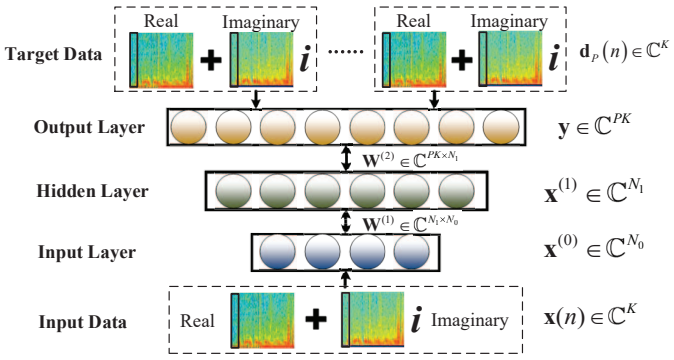

Fig. 1. The architecture of FCDNN for source separation.

3. FCDNN-BASED SOURCE SEPARATION

Unlike previously developed DNN-based methods, the proposed method directly estimates the complex-valued STFT coefficient of each source. To preserve the structure of complex-valued STFT coefficients, the FCDNN is developed to learn the nonlinear relationship between the input mixture and the target sources in the complex domain. The concept of complex-valued neural network (CVNN) was originally proposed in [17] and [18]. The FCDNN that is developed in this paper is a deeper version of the CVNN. Fig. 1 displays an overview of FCDNN-based source separation. Specially, the FCDNN operates directly in the complex domain. The activations and weights of the FCDNN, which can be regarded as the spectral pattern of STFT coefficients, are all complex-valued. Notably, a deeper CVNN has also been developed to perform a beamforming task [19]. However, its object function is unconstrained, potentially resulting in an overfitting model in the complex domain.

3.1. Sparse Model Training

Herein, the goal is to learn a nonlinear mapping from X to S_p in the complex domain. To train the FCDNN, the STFT coefficients of the mixture are concatenated into a complex-valued feature vector for frame n, which is defined as

$$\mathbf{x}(n) = (X(n,1), X(n,2), ..., X(n,K)) \in \mathbb{C}^K, 1 \leq n \leq N \tag{6}$$

where K is the number of frequency bins, and N is the number of frames. Similarly, the STFT coefficients of the p-th source for the n-th frame can be represented by

$$\mathbf{d}_p(n) = (S_p(n,1), S_p(n,2), ..., S_p(n,K)) \in \mathbb{C}^K \tag{7}$$

Given the pair of data $\{\mathbf{x}(n), \mathbf{d}_p(n)\}$, the objective function of the FCDNN can be defined as follows,

$$E = \sum_{n=1}^{N} E_n = \sum_{n=1}^{N} \left(\mathbf{d}(n) - \mathbf{y}(n)\right)\left(\mathbf{d}(n) - \mathbf{y}(n)\right)^H \in \mathbb{R} \tag{8}$$

where $\mathbf{y}(n) \in \mathbb{C}^{KP}$ is the output of the FCDNN; E_n is the n-th partial error term, $\mathbf{d}(n) = (\mathbf{d}_1(n), \mathbf{d}_2(n), ..., \mathbf{d}_P(n)) \in \mathbb{C}^{KP}$ is the concatenation of the P sources, and H is the Hermitian transpose.

Without loss of generality, a two-layer FCDNN is considered, as shown in Fig. 1. Omitting the frame index n, the j-th element of can be represented as

$$y_j = x_j^{(2)} = f(\underbrace{\sum_{k=1}^{N_1} w_{jk}^{(2)} \cdot f\left(a_k^{(1)}\right) + b_j^{(2)}}_{a_j^{(2)}}) \in \mathbb{C} \qquad (9)$$

where $a_k^{(1)} = \sum_{m=1}^{N_0} w_{km}^{(1)} x_m^{(0)} + b_k^{(1)}$; $f : \mathbb{C} \to \mathbb{C}$ is a nonlinear activation function in the complex domain; $x_k^{(0)} \in \mathbb{C}$ denotes the k-element of the inputs, $x_j^{(l)} \in \mathbb{C}$ represents the j-th network outputs, $w_{jk}^{(l)} \in \mathbb{C}$ are the network weights, and $b_j^{(l)} \in \mathbb{C}$ are the j-th network biases for $l \in \{1, 2\}$. Notably, all of the network parameters are complex-valued.

The discriminative term [6, 8–10] is commonly incorporated into the objective function to regularize the reconstruction error. Unlike such approaches, the method in this work considers prior knowledge of the inherent sparse structure of speech signals in the time-frequency domain. Motivated by the sparse auto-encoder [20], a sparse constraint is further imposed on the objective function of the FCDNN. Instead of applying the batch objective function in Eq. (8), the object function which is calculated by using n-th sample $\mathbf{x}(n)$ is considered:

$$E_n^{\text{sparse}} = E_n + \beta \cdot \sum_{j=1}^{M} D_{\text{KL}}\left(\rho \parallel \hat{\rho}_{nj}\right) \qquad (10)$$

where $\hat{\rho_j} = \frac{1}{m} \sum_{i=1}^{m} \left| f(a_j^{(l)}) \right|$ denotes the mean activation of the j-th hidden unit; $a_j^{(l)}$ is the j-th activation in l-layer; M represents the number of neurons in the l-layer; ρ is the predefined sparse parameter, and β controls the balance between the error term and the sparse penalty term. In this paper, l was set to specify the last layer of the FCDNN.

The first term on the right-hand side of Eq. (10) can be interpreted as a reconstruction error term and the second term can be interpreted as a sparse penalty term. The sparse term, $D_{\text{KL}}\left(\rho \parallel \hat{\rho}_{nj}\right)$ is the KL divergence between the mean activation and the pre-defined sparsity that forces a large number of time-frequency bins to be "inactive" in the estimated spectra. The advantage is that the number of free parameters of the FCDNN is reduced, ensuring that the model does not find a poor local minimum during the learning.

3.2. Complex-valued Activation

With respect to the activation function of the FCDNN, the Liouville theorem [21] states that every bounded function is constant in the complex domain. Therefore, activation functions that are typically used in real-valued DNNs are not suitable for use with the FCDNN. Moreover, the learning problem in a complex domain is more complicated than that in a real domain. Based on the impressive result that the rectified linear unit (ReLU) is easier to learn than conventional activations [22], a complex-valued ReLU [23] is adopted in the FCDNN in this paper, and is defined as,

$$\text{ReLU}_{\mathbb{C}}(z) = \begin{cases} z & ,\phi_z \in \left[0, \frac{\pi}{2}\right] \\ 0 & ,\text{otherwise} \end{cases} \qquad (11)$$

The $\text{ReLU}_{\mathbb{C}}$ is found herein to be less sensitive to the initialization of weights than other complex-valued activations, such as tanh and sigmoid, in the source separation task.

3.3. Error Back-propagation for FCDNN

To train the developed FCDNN for source separation, the stochastic gradient decent (SGD) is adopted in our work. Other state-of-the-art back-propagation methods are evaluated in this work, such as the Quasi-Newton method [24], but they were not as effective as SGD. To update the network parameters using SGD, the gradient of the n-partial error term E_n^{sparse} is required. Then, the network parameters can be updated based on one data point:

$$\theta^{(\tau+1)} = \theta^{(\tau)} + \Delta\theta^{(\tau)} = \theta^{(\tau)} - \eta \frac{\partial E_n^{\text{sparse}}}{\partial \theta^{(\tau)}} \qquad (12)$$

where $\theta = \left\{ w_{jk}^{(l)}, b_k^{(l)} \,|\forall j, k, l \right\}$ is the concatenation of all network parameters; η is the learning late, and τ is the iteration number. $\mathbb{CR}$-calculus [25] is utilized to calculate the partial derivative of E_n^{sparse} with respect to complex-valued parameters. For example, the partial derivative of E_n^{sparse} with respect to $w_{jk}^{(2)}$ can be calculated by,

$$\frac{\partial E_n^{\text{sparse}}}{\partial (w_{jk}^{(2)})} = \frac{\partial E_n}{\partial (w_{jk}^{(2)})^{\Re}} + i \cdot \frac{\partial E_n}{\partial (w_{jk}^{(2)})^{\Im}} + \beta \cdot \left(-\frac{\rho}{\hat{\rho}_{nk}} + \frac{1-\rho}{1-\hat{\rho}_{nk}} \right) \cdot x_k^{*(1)} \qquad (13)$$

Specially, the first term in the right-hand side of Eq. (13) is derived as

$$\begin{aligned} \frac{\partial E_n}{\partial (w_{jk}^{(2)})^{\Re}} &= \frac{\partial E_n}{\partial \mathrm{y}_j^*} \cdot \frac{\partial \mathrm{y}_j^*}{\partial a_j^{*(2)}} \cdot \frac{\partial a_j^{*(2)}}{\partial (w_{jk}^{(2)})^{\Re}} + \frac{\partial E_n}{\partial \mathrm{y}_j} \cdot \frac{\partial \mathrm{y}_j}{\partial a_j^{(2)}} \cdot \frac{\partial a_j^{(2)}}{\partial (w_{jk}^{(2)})^{\Re}} \\ &= -\left(d_j - \mathrm{y}_j\right) f'\left(a_j^{*(2)}\right) x_k^{*(1)} - \left(d_j^* - \mathrm{y}_j^*\right) f'\left(a_j^{(2)}\right) x_k^{(1)} \end{aligned} \qquad (14)$$

where $a_j^{(2)} = w_{jk}^{(2)} x_k^{(1)} + b_j^{(2)}$; $*$ indicates the complex conjugate operations; $f' : \mathbb{C} \to \mathbb{C}$ is the first-order derivative of the activation function. Similarly, the second term in Eq. (13) can be obtained:

$$\frac{\partial E_n}{\partial (w_{jk}^{(2)})^{\Im}} = \left(d_j - \mathrm{y}_j\right) f'\left(a_j^{*(2)}\right)\left(i \cdot x_k^{*(1)}\right) - \left(d_j^* - \mathrm{y}_j^*\right) f'\left(a_j^{(2)}\right)\left(i \cdot x_k^{(1)}\right) \qquad (15)$$

Finally, given the STFT coefficients of the testing samples, the STFT cofficients of target sources can be estimated by the forward pass of FCDNN.

4. EXPERIMENTS

The effectiveness of the proposed method is experimentally evaluated on the singing source separation task. The performance of source separation that is evaluated using BSS-EVAL metrics [26], including SIR, SAR, and SDR. Only the vocals are evaluated. SNR$_{fw}$ [27] and PESQ [28] were used to measure the quality of separated vocals. In the experiment, 1000 song clips from the MIR-1K database are used [29]. Each song involves two sources: one is the vocals and the other is the instrumental music. All of the sound clips are recorded at a sampling rate of 16 kHz and are between 4 and 13s long. To generate the training and development set, 175 clips of songs are selected from MIR-1K. For the testing set, the remaining 825 clips of songs are used. Two sources ($P = 2$) are mixed to form the mixture with equal energy.

The spectrograms were generated using a 128-point STFT with Hamming windows ($K = 65$). The windows were shifted relative to each other by one half of the window length so that they overlapped. The spectra of the mixed clips are combined with those of the preceding and following five frames, and then used as the input in the DNN models. This gave approximately 347715 paired samples ($N = 347715$) for training. The architecture of the DNN models is fixed to 715-2500-2500-130, indicating that the sizes of the input layer, the two hidden layers, and the two source signals in the output layer were 715 (65×11), 2500 and 130 (65×2), respectively. The highest epoch is set to 200. The learning rates were set to 0.001, 0.001 and 0.0001 for the input-hidden neurons, the hidden-hidden neurons and the target-hidden neurons, respectively. For the sparse-constrained FCDNN (FCDNN-S), the β and ρ were empirically set to 0.005 and 10^{-8}, respectively.

Table 1. Performance comparison between proposed methods and baseline methods in terms of SNR$_{fw}$ and PESQ.

Methods	SNR$_{fw}$	PESQ
Mixture	-0.89±1.29	1.22±0.43
IRM	5.36±1.37	1.99±0.41
DNN-M	0.56±1.66	1.45±0.37
DNN-RI	1.65±2.00	1.53±0.33
FCDNN	1.50±1.90	1.50±0.34
FCDNN-S	1.83±2.02	1.59±0.33

To confirm the efficiency of the FCDNN-based methods, a standard DNN-based method: DNN-M [7], which mainly involve modification of the spectra magnitude, is selected as the baseline. Since the DNN-M improve only the magnitude response of spectra, another state-of-the-art method [14] (DNN-RI), which jointly estimates the real and imaginary components, is also compared to the proposed FCDNN. The size of the input and output layers in DNN-RI was twice that in

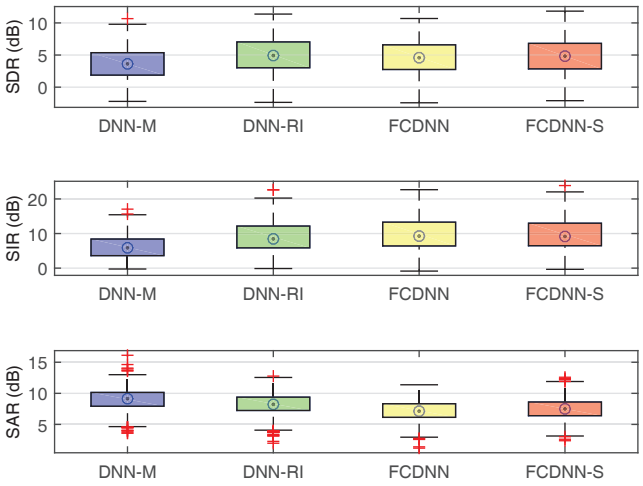

Fig. 2. Results of source separation using the MIR-1K dataset by the baseline methods and the FCDNN-based methods.

FCDNN. To ensure a fair comparison, the proposed method and the baseline methods use the STFT coefficients as time-frequency features. The DNN-based methods adopt ReLU as the activation function. Moreover, all methods use apply the standard SGD to optimize the network parameters. Fig. 2 shows the experimental results in terms of SIR, SAR and SDR. Experimental results demonstrate that the proposed method outperformed the baseline methods in terms of SDR and SIR. However, FCDNN achieved lower SAR compared with the baseline methods. Table 1 shows the average performance in terms of SNR$_{fw}$ and PESQ. FCDNN had a better PESQ than DNN-M, but its PESQ was similar to that of DNN-RI. Comparison between FCDNN and FCDNN-S confirmed the power of the additional sparse regularization term.

5. CONCLUSIONS AND FUTURE WORK

This work presented a novel FCDNN-based method for monaural source separation. To incorporate the phase information, which ignored by the majority of source separation approaches, the developed FCDNN is employed to learn the nonlinear mapping between the input mixture and the target sources. Unlike conventional DNN-based methods, the proposed method operates directly in the complex domain, and also provides an intuitive way to deal with complex-valued signals. Additionally, a sparsity constraint is imposed on the objective function of FCDNN, enforcing the regularity of the learned model. Experimental results indicate that the proposed method has significantly higher SDR and SIR than two state-of-the-art methods. Moreover, the proposed method yields better performance on the perceptual quality than the conventional DNN-based method.

6. REFERENCES

[1] H. B. Barlow, *Underlying the Transformations Sensory Messages*, Cambridge, MA: MIT Press, 1961.

[2] N. Cho and C.-C. J. Kuo, "Sparse music representation with source-specific dictionaries and its application to signal separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 2, pp. 326–337, 2011.

[3] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. ICA'07*, 2007, pp. 414–421.

[4] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, and K. Kondo, "Robust music signal separation based on supervised nonnegative matrix factorization with prevention of basis sharing," in *Proc. ISSPIC*, 2013, pp. 392–397.

[5] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, "Nmf-based target source separation using deep neural network," *IEEE Signal Process. Lett.*, vol. 22, no. 2, pp. 229–233, 2015.

[6] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. ICASSP*, 2014, pp. 1562–1566.

[7] Y. X. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[8] G. X. Wang, C. C. Hsu, and J. T. Chien, "Discriminative deep recurrent neural networks for monaural speech separation," in *Proc. ICASSP*, 2016, pp. 2544–2548.

[9] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 12, pp. 1–12, 2015.

[10] S. Nie, S. Liang, H. Li, X. L. Zhang, Z. L. Yang, W. J. Liu, and L. K. Dong, "Exploiting spectro-temporal structures using nmf for dnn-based supervised speech separation," in *Proc. ICASSP*, 2016, pp. 469–473.

[11] K. Paliwal, K. Wojcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, 2011.

[12] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, 2015.

[13] M. Krawczyk and T. Gerkmann, "Stft phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1931–1940, 2014.

[14] D. S. Williamson, Y. X. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 3, pp. 483–492, 2016.

[15] D. S. Williamson, Y. X. Wang, and D. L. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *Proc. ICASSP*, 2016.

[16] A. J. R. Simpson, "Deep transform: Cocktail party source separation via complex convolution in a deep neural network," *arXiv preprint arXiv:1504.02945*, 2015.

[17] H. Leung and S. Haykin, "The complex backpropagation algorithm," *IEEE Trans. Signal Process.*, vol. 39, no. 9, pp. 2101–2104, 1991.

[18] N. Benvenuto and F. Piazza, "On the complex backpropagation algorithm," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 967–969, 1992.

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

[20] A. Ng, "Sparse autoencoder," Tech. Rep., CS294A Lecture Notes, Stanford Univ., CA, USA, 2011.

[21] M. J. Ablowitz and A. S. Fokas, *Complex Variables*, Cambridge, U.K.: Cambridge Univ. Press, 2003.

[22] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, 2012.

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

[24] J. Sohl-Dickstein, B. Poole, and S. Ganguli, "Fast large-scale optimization by unifying stochastic gradient and quasi-newton methods," in *Proc. ICML*, 2014.

[25] K. Kreutz-Delgado, "The complex gradient operator and the cr-calculus. univ," Tech. Rep., Univ. California, San Diego, 2009.

[26] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[27] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.

[28] M. P. Hollier A. W. Rix, J. G. Beerends 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.

[29] C. L. Hsu and J. S. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 310–319, 2010.