CROSS-DOMAIN SPEECH ENHANCEMENT WITH A NEURAL CASCADE ARCHITECTURE

Heming Wang[1] *and DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA
wang.11401@osu.edu, dwang@cse.ohio-state.edu

ABSTRACT

This paper proposes a novel cascade architecture to address the monaural speech enhancement problem. We leverage three different domains of speech representation, namely spectral magnitude, waveform, and complex spectrogram, to progressively suppress the background noise within noisy speech. Our proposed neural cascade architecture consists of three modules, and each operates on the original noisy input and the output of the previous module in a distinct speech representation. During training, the network simultaneously optimizes all modules with a triple-domain loss. Experiments on the WSJ0 SI-84 corpus demonstrate that our proposed approach achieves superior enhancement results, and substantially outperforms previous baselines in terms of both speech quality and intelligibility.

Index Terms— speech enhancement, spectral magnitude, time domain, complex domain, cross-domain speech enhancement

1. INTRODUCTION

Background noise is an unavoidable interference in real-world speech communication, and is harmful for speech processing tasks like automatic speech recognition. The goal of speech enhancement is to remove background noise and recover the clean speech signal, in order to improve the speech quality and intelligibility of noisy speech. Conventional methods include traditional enhancement techniques like spectral subtraction [1] and computational auditory scene analysis [2]. Since the introduction of deep learning, dramatic progress has been made in this field [3].

Early deep neural network (DNN) based studies use spectral magnitude features as training targets. Those include the ideal binary mask [4], ideal ratio mask (IRM) [5], and target magnitude spectrum [6]. Recent studies address phase enhancement [7], as speech phase proves to be important for speech quality. To address phase estimation, one popular direction is to estimate the real and imaginary parts of the complex spectrogram of clean speech [8, 9, 10]. The other popular approach is to estimate the clean speech waveform in the time domain [11, 12, 13].

Although these approaches are effective for noise suppression, they are single-stage models that utilize only one speech representation. Recent studies investigate combining different speech representations for improved speech enhancement. Some attempt to incorporate a different representation domain into the loss function. For example, Pandey and Wang [12] train a time-domain DNN, but optimize the network in the frequency domain by performing short-time Fourier transform (STFT) on the predicted signals and clean signals. Bahmaninezhad et al. [14] feed an DNN with frequency-domain features. During training, a scale-invariant signal-to-noise ratio (SNR) based loss is calculated by converting predicted complex vectors to the time domain with inverse STFT (iSTFT). Others attempt to enhance speech in multiple stages and each stage operates in one signal domain. Tzinis et al. [15] propose a two-step architecture for source separation. Instead of directly separating the sources, they first learn a latent speech representation, and then perform speech separation in the learned latent embedding space. Lin et al. [16] stack temporal convolutional modules to progressively enhance speech magnitude in multiple stages. Each module (except the first module) takes as input the noisy magnitude and the predicted output from the last module. Li et al. [17] introduce a two-stage network where the first stage only estimates the magnitude, and then the second stage performs complex spectral mapping using the predicted magnitude spectrum and the original noisy spectrum. Zhang et al. [18] propose a two-stage framework that incorporates multiple training targets. In the first stage, they perform joint training using two branches to predict the complex spectrum and IRM, respectively. In the next stage, they use the enhanced magnitude obtained from the first stage to predict the prior SNR.

In this paper, we propose a novel neural cascade architecture (NCA) that combines the strengths of cross-domain speech representations. NCA consists of three modules that operate on the spectral magnitude, waveform and complex spectrogram, respectively. Each module operates on the out-

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

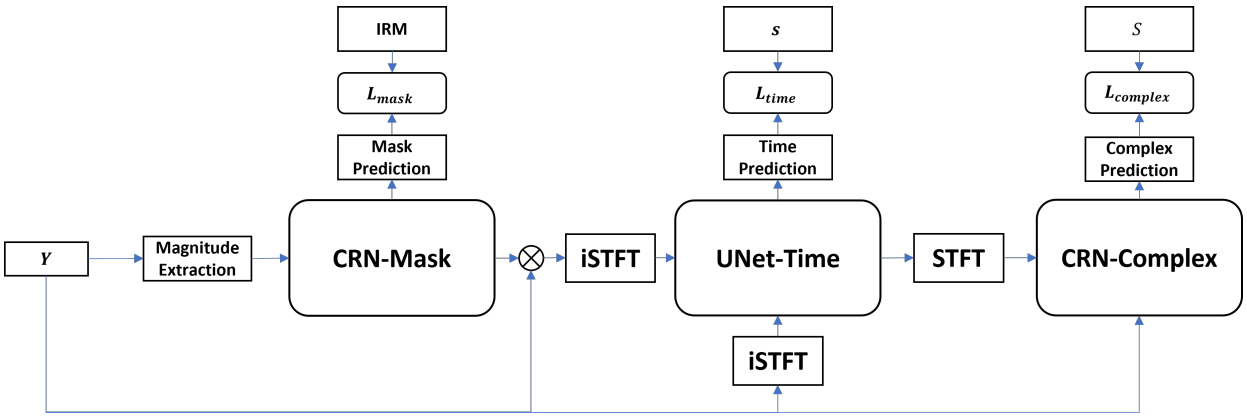


Fig. 1. Diagram of the cascade architecture. The input Y is the complex spectrogram of the noisy input. Three outputs are produced by the network, which are the mask, the time and the complex predictions.

put of the previous module and the original noisy input. We optimize all modules simultaneously with a triple-domain loss. Different from other studies that incorporate one or two domains of speech representation, our network incorporates three training targets in DNN-based speech enhancement. In addition, our approach is trained in an end-to-end fashion, which saves training effort, whereas other multi-stage models typically employ complicated training strategies such as pre-training and fine-tuning. Experimental results show that our model achieves significantly better enhancement performance compared to previous strong baselines.

2. NEURAL CASCADE ARCHITECTURE

As illustrated Fig. 1, NCA is composed of three modules, CRN-Mask, UNet-Time and CRN-Complex. As the names suggest, the three modules employ the popular design of recent enhancement studies, which is elaborated in Section. 2.1. CRN-Mask is fed with magnitude features and predicts the IRM [4]. The next module UNet-Time accepts two time-domain inputs, in which one is converted from the masked spectrogram from the previous module by taking the iSTFT, and the other is the original noisy waveform. Such design mitigates the estimation error and distortion brought by the previous module. Similarly, the last module CRN-Complex takes in as input the noisy complex spectrogram and the output of UNet-Time. Given a noisy speech mixture y that is composed of background noise n and clean speech s, with the parameters of each module denoted as θ, our pipeline can be formulated as,

$$\begin{aligned}\hat{S}_1(t,f) &= f_{mask}(\theta_{mask}, |Y(t,f)|) \odot Y(t,f) \\ \hat{s}_2(k) &= f_{time}(\theta_{time}, y(k), \hat{s}_1(k)) \\ \hat{S}_3(t,f) &= f_{complex}(\theta_{complex}, Y(t,f), \hat{S}_2(t,f)),\end{aligned} \quad (1)$$

where the subscript number $1, 2, 3$ corresponds to each module, k indicates a time sample, and $\odot$ indicates the element-wise multiplication. The capital letters S, Y are the STFTs of s and y, and t and f index the time step and frequency bin, respectively. We regard $\hat{S}_3$ as our final enhancement result.

2.1. Module Design

Fig. 2 depicts the detailed design of modules within NCA. Both the mask and complex modules are based on the convolutional recurrent network (CRN) architecture [9]. CRN is composed of an encoder, a recurrent neural network based bottleneck, and a decoder. We employ a pointwise convolution for skip connections and a two-layer grouped LSTM as our bottleneck. For CRN-Mask, we append a linear layer with a sigmoidal activation function after the regular CRN to produce the mask prediction. For CRN-Complex, we replace each CNN layer with a densely connected (DC) block [19]. The DC block consists of 5 convolutional layers with a growth rate of 8, and within the block all layers are directly connected. Moreover, we split the CRN output into two halves and each is followed by a linear layer to predict real and imaginary parts separately. We employ the encoder-decoder structure of the standard UNet [20] for the time module. UNet-Time performs frame-level speech enhancement and is fed with speech segments.

2.2. Triple-domain Loss

We propose a novel triple-domain loss as the training objective. As described in the previous sections, we obtain a IRM estimation RM, and time domain estimate $\hat{s}$ and a complex spectrum prediction $\hat{S}$ from all modules. IRM is calculated based on the energy of speech S and noise N in the time-frequency (T-F) domain, defined as $\mathrm{IRM}(t,f) = \sqrt{\frac{S(t,f)^2}{S(t,f)^2+N(t,f)^2}}$. L_{mask} measures the mean absolute error of the mask estimation and the ground truth IRM.

$$L_{mask} = \frac{1}{TF}\sum_{t,f} |\mathrm{RM}(t,f) - \mathrm{IRM}(t,f)|, \quad (2)$$

where T and F are the total number of time steps and frequency bins. L_{time} employs the phase-constrained magnitude loss proposed in [13], as experiments demonstrate that it effectively imposes a phase constraint and leads to a good magnitude estimate.

$$L_{time} = \frac{1}{TF}\sum_{t,f}[(|\hat{S}(t,f)| - |S(t,f)|) + (|Y(t,f) - \hat{S}(t,f)| - |N(t,f)|)]. \quad (3)$$

$L_{complex}$ measures the difference of complex representations as well as the magnitude difference, as it has been addressed in previous literature [21, 22] that a good magnitude estimation considerably boosts the enhancement performance in complex spectral mapping,

$$L_{complex} = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}[||\hat{S}(t,f)| - |S(t,f)|| + (|\hat{S}_r(t,f) - S_r(t,f)| + |\hat{S}_i(t,f) - S_i(t,f)|)]. \quad (4)$$

Here the subscripts r and i represent the real and imaginary parts of T-F representations, respectively.

The triple-domain loss is defined upon these three loss functions,

$$L_{triple} = \lambda_1 L_{mask} + \lambda_2 L_{time} + L_{complex}. \quad (5)$$

We assign $\lambda_1 = 5.0$ and $\lambda_2 = 1.0$ to balance different value ranges of the three loss terms.

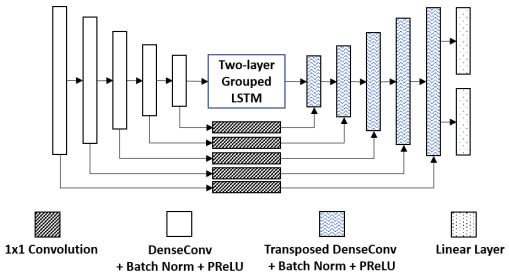
3. EXPERIMENT

3.1. Training dataset

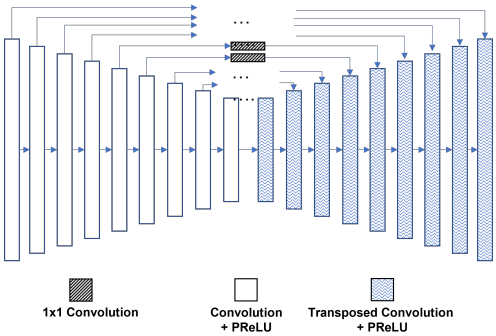
We conduct experiments on the WSJ0 SI-84 dataset [23], which contains English utterances uttered by 42 male and 41 female speakers. Of the 7138 utterances within the dataset, we select 5428 utterances from 77 speakers to generate the training set. Noise files in the DNS Challenge[1] are utilized, and we randomly pick 20000 noise files of a total duration of 55 hours to construct the training noise. During training data generation, for each clean utterance, we randomly cut a segment from the training noise that is the same length and then mix them at a SNR uniformly sampled from -5, -4, -3, -2, -1, and 0 dB. Repeating that procedure, we create 50000 mixtures as the training set. Using 150 clean utterances selected outside the training scope, we follow a similar process to create the validation set of 4000 mixtures. For evaluation, we choose 6 untrained speakers, each having 25 utterances. Four challenging noises are utilized, which are babble (denoted as babble1), factory1 from NOISEX92 [24], and babble (denoted as babble2) and cafeteria from an Auditec CD[2]. The testing mixtures are generated by mixing with these noises at three different SNR levels -5, 0 and 5 dB.

[1]https://github.com/microsoft/DNS-Challenge

[2]available at http://www.auditec.com



(a) CRN-Complex



(b) UNet-Time

Fig. 2. Illustration of the modules of the NCA. From top to bottom: (a). The complex module CRN-Complex, (b). The time-module UNet-Time. CRN-Mask and CRN-Complex are similar, so we just display the diagram of the complex module for brevity.

3.2. Experimental Settings

All the utterances in the experiment are sampled at 16 kHz. We select the Hamming window with a window length of 320 samples and a window shift of 160 samples during STFT operation. During training, we use the Adam optimizer to perform stochastic gradient descent optimization. The training epoch is set to be 50, and we use a batch size of 8 utterances with an initial learning rate of 0.001. The enhancement performance is measured by two metrics, perceptual evaluation of speech quality [25] and extended short-term objective ineligibility (ESTOI) [26]. PESQ evaluates the speech quality and has a value ranges from -0.5 to 4.5. ESTOI evaluates speech intelligibility with a value within 0 and 1 and can be interpreted as the percentage of correctness. For both metrics, higher values suggest better enhancement performance.

3.3. Experimental Results

We compare our proposed NCA with four strong speech enhancement baselines in both causal and non-causal settings. Those include complex-domain approaches, gated CRN (GCRN) [27] and deep complex CRN [10] (DCCRN). Moreover, we compare with the time-domain baseline autoencoder CNN (AECNN) [12], and the two-stage baseline Complex spectral mapping based Two-Stage Network (CTSNet) [17].

First, we present in Table 1 and 2 the evaluation results

Table 1. Evaluations and Comparisons of Different Enhancement Models in terms of ESTOI(%)

	SNR	-5 dB					0 dB					5 dB				
	Causal	Babble1	Factory	Babble2	Cafeteria	Average	Babble1	Factory	Babble2	Cafeteria	Average	Babble1	Factory	Babble2	Cafeteria	Average
Unprocessed	-	26.40	26.75	25.62	24.25	25.76	39.47	40.94	38.81	37.92	39.29	54.05	56.77	53.89	53.44	54.54
GCRN	✓	51.10	52.56	51.40	48.40	50.87	68.72	69.17	69.87	67.05	68.71	80.45	80.68	81.74	79.45	80.58
AECNN	✓	57.88	59.72	48.87	47.81	53.57	71.87	72.21	66.02	64.97	68.77	79.62	79.71	79.12	78.18	79.16
DCCRN	✓	54.72	55.41	54.78	52.39	54.33	71.71	71.59	72.68	70.38	71.59	82.66	82.59	83.89	81.94	82.77
CTSNet	✓	60.06	61.05	60.21	56.43	59.44	75.86	75.10	74.74	74.03	74.93	84.54	83.89	86.44	84.48	84.84
NCA	✓	**63.96**	**63.41**	**65.76**	**60.40**	**63.38**	**78.83**	**77.75**	**80.56**	**76.89**	**78.51**	**86.50**	**85.82**	**87.73**	**85.37**	**86.36**
BGCRN	✗	56.83	58.79	57.34	54.71	56.92	73.52	73.79	75.00	72.31	73.66	83.66	83.64	84.64	82.72	83.67
BDCCRN	✗	57.46	58.79	58.26	55.78	57.57	74.76	74.25	75.65	73.13	74.45	84.50	84.11	85.56	83.54	84.43
NC-CTSNet	✗	63.01	63.42	63.09	60.24	62.44	78.91	77.51	79.91	76.94	78.32	86.19	85.70	87.47	85.48	86.21
NC-NCA	✗	**69.06**	**68.29**	**70.61**	**65.26**	**68.31**	**81.71**	**80.55**	**83.54**	**80.00**	**81.45**	**88.10**	**87.32**	**89.23**	**87.15**	**87.95**

Table 2. Evaluations and Comparisons of Different Enhancement Models in terms of PESQ

	SNR	-5 dB					0 dB					5 dB				
	Causal	Babble1	Factory	Babble2	Cafeteria	Average	Babble1	Factory	Babble2	Cafeteria	Average	Babble1	Factory	Babble2	Cafeteria	Average
Unprocessed	-	1.54	1.44	1.55	1.46	1.50	1.83	1.75	1.82	1.77	1.79	2.15	2.10	2.11	2.13	2.12
GCRN	✓	1.71	1.95	1.73	1.75	1.79	2.37	2.56	2.43	2.43	2.45	2.94	3.03	2.98	2.93	2.97
AECNN	✓	2.10	2.31	1.82	1.95	2.05	2.68	2.36	2.54	2.43	2.50	2.90	2.92	2.89	2.91	2.91
DCCRN	✓	1.98	2.08	1.93	1.98	1.99	2.49	2.52	2.52	2.49	2.51	2.92	2.89	2.97	2.91	2.92
CTSNet	✓	2.12	2.29	2.13	2.11	2.16	2.74	2.78	2.67	2.78	2.74	3.21	3.18	3.22	3.17	3.20
NCA	✓	**2.15**	**2.35**	**2.22**	**2.20**	**2.23**	**2.81**	**2.88**	**2.91**	**2.83**	**2.86**	**3.27**	**3.26**	**3.34**	**3.23**	**3.28**
BGCRN	✗	2.08	2.31	2.06	2.10	2.14	2.31	2.75	2.82	2.74	2.66	3.20	3.23	3.19	3.18	3.20
BDCCRN	✗	2.05	2.23	2.06	2.11	2.11	2.65	2.68	2.67	2.62	2.66	3.05	3.02	3.08	3.01	3.04
NC-CTSNet	✗	2.19	2.39	2.20	2.23	2.25	2.90	2.91	2.93	2.85	2.90	3.32	3.29	3.33	3.27	3.30
NC-NCA	✗	**2.43**	**2.54**	**2.49**	**2.40**	**2.47**	**3.05**	**3.05**	**3.13**	**3.02**	**3.06**	**3.43**	**3.40**	**3.47**	**3.40**	**3.43**

Table 3. Effects of Different Optimization Strategies at -5 dB SNR.

Model	ESTOI(%)	PESQ	Training time
End-to-end optimization	**63.38**	**2.23**	1.0x
Multi-stage sequential training	60.81	2.11	1.9x
Multi-stage joint training	63.09	**2.23**	2.1x
Only optimizing $L_{complex}$	58.56	2.09	1.0x

of causal networks on the WSJ0 SI-84 corpus. As shown in the table, our proposed network performs the best under all conditions. In addition, the multi-stage approaches CTSNet and NCA considerably outperform single-stage paradigms GCRN, DCCRN and AECNN in terms of both PESQ and ESTOI. Moreover, the proposed NCA consistently outperforms the strongest baseline CTSNet. For example, at -5 dB SNR, the ESTOI is improved by 3.94%, and PESQ by 0.07 on average.

We also convert the causal models to their non-causal versions (denoted as BDCCRN, BGCRN, NC-CSTNet and NC-NCA), and present their results in the tables. In baseline models, all causal convolutions are replaced with non-causal convolutions and we replace LSTM layers with bidirectional LSTMs. The conversion for non-causal AECNN is not straightforward, so we do not include it in the tables. Non-causal networks show a significant performance improvement compared to their causal counterparts, as they utilize future information. The performance advantage of NC-NCA is maintained, and the gap with the best baseline model NC-CTSNet is even larger. For instance, the PESQ is improved by 0.22 and ESTOI is improved by 5.87% on average at -5 dB SNR.

We further analyze the effect of various optimization strategies for the NCA in Table 3. We use the result of causal NCA under -5 dB as the baseline and compare the enhancement performance under the same settings with other optimization strategies. First, we sequentially train the three modules within the NCA. CRN-Mask is trained first, and then with the first module frozen, we train UNet-Time with the output of the first module and the noisy input. Finally, the first two modules are frozen, and we train CRN-Complex using the original input and the prediction obtained from UNet-Time. We also investigate the joint training approach, which is similar to the sequential training except we do not freeze any modules. Instead, for the second and last step, we jointly train the current module with the previous module(s), and set a smaller learning rate for the previous module(s). Finally, we explore training the NCA with only the complex loss $L_{complex}$ measured at the last module. As shown in the table, the end-to-end training strategy has the overall best performance. Multi-stage joint training also produces excellent performance, but it is not preferable as it requires around twice the training time and a more sophisticated training strategy. Only optimizing $L_{complex}$ dramatically degrades the enhancement performance, suggesting that it is crucial to impose optimization constraints on all modules. Otherwise, the first two modules could not produce outputs beneficial for the following module(s).

4. CONCLUSION

We have proposed a novel neural cascade architecture to leverage the strengths of three different speech representations, which are, magnitude, waveform and complex spectrogram. Our model is trained end-to-end with a triple-domain loss. Experiments have demonstrated the superiority of our design, and we achieve significantly better enhancement results compared with other advanced baselines. Future work includes a faster and lighter model design for real-time mobile applications. In addition, we plan to investigate cross-domain features to further improve generalization.

5. REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[2] D. L. Wang and G. J. Brown Eds, *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley-IEEE press, 2006.

[3] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1702–1726, 2018.

[4] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, pp. 181–197. 2005.

[5] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1849–1858, 2014.

[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, pp. 65–68, 2014.

[7] P. Mowlaee, R. Saeidi, and R. Martin, "Phase estimation for signal reconstruction in single-channel source separation," in *Proceedings of INTERSPEECH*, 2012, pp. 1548–1551.

[8] 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, pp. 483–492, 2016.

[9] K. Tan and D. L. 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, 2020.

[10] 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 *Proceedings of INTERSPEECH*, 2020, pp. 2482–2486.

[11] C. Macartney and T. Weyde, "Improved speech enhancement with the Wave-U-Net," *arXiv:1811.11307*, 2018.

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

[13] A. Pandey and D. L. Wang, "Dense CNN with self-attention for time-domain speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1270–1279, 2021.

[14] 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 *Proceedings of INTERSPEECH*, 2019, pp. 4574–4578.

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

[16] J. Lin, A. J. van Wijngaarden, K.-C. Wang, and M. C. Smith, "Speech enhancement using multi-stage self-attentive temporal convolutional networks," *arXiv:2102.12078*, 2021.

[17] 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.

[18] L. Zhang, M. Wang, A. Li, Z. Zhang, and X. Zhuang, "Incorporating multi-target in multi-stage speech enhancement model for better generalization," *arXiv:2107.04232*, 2021.

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

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

[21] Z.-Q. Wang and D. L. Wang, "Multi-microphone complex spectral mapping for speech dereverberation," in *Proceedings of ICASSP*, 2020, pp. 486–490.

[22] Z.-Q. Wang, G. Wichern, and J. L. Roux, "On the compensation between magnitude and phase in speech separation," *IEEE Signal Processing Letters*, 2021.

[23] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of a Workshop on Speech and Natural Language*, 1992.

[24] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, 1993.

[25] 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 *Proceedings of ICASSP*, 2001, vol. 2, pp. 749–752.

[26] J. Jensen and C. H. 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. 2009–2022, 2016.

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