A Speech Enhancement Method Combining Two-Branch Communication and Spectral Subtraction

Ruhan He[1,2], Yajun Tian[1,2](✉), Yongsheng Yu[3], Zhenghao Chang[1,2], and Mingfu Xiong[1,2]

[1] Hubei Provincial Engineering Research Center for Intelligent Textile and Fashion, Wuhan 430200, China
1316108825@qq.com

[2] School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, China

[3] State Key Laboratory of Silicate Materials for Architectures, Wuhan University of Technology, Wuhan 430070, China

Abstract. Time-Frequency (T-F) domain masking is currently the dominant method for single-channel speech enhancement, while little attention has been paid to phase information. A speech enhancement method, named PHASEN-SS, is proposed in this paper. Our method is divided into two steps, first a deep neural network (DNN) with two-branch communication using a combination of mask and phase for speech enhancement, and then a data post-processing after the DNN processes the noisy speech. PHASEN-SS uses two branches to predict the amplitude mask and the phase separately, which improves the accuracy of prediction by exchanging information between two branches, and then further the enhancement by denoising the residual noise through spectral subtraction. The experiments are conducted on the publicly available Voice Bank + DEMAND dataset, as well as a noisy speech dataset is synthesized with 4 common noises in Noise92 and Voice Bank clean speech according to the specified signal-to-noise ratio (SNR). The results show that the proposed method improves on the original one, and has better robustness to speech containing babble noise at higher SNRs for different SNRs.

Keywords: spectral subtraction · neural network · phase prediction · speech enhancement · time-frequency mask

1 Introduction

There are two main types of speech enhancement methods: traditional methods and deep learning methods. Among the traditional methods, speech enhancement methods include spectral subtraction [1], Wiener filtering [2], statistical model-based methods [3] and subspace algorithms [4,5], etc., which are more

M. Tanveer et al. (Eds.): ICONIP 2022, CCIS 1792, pp. 110–122, 2023.
https://doi.org/10.1007/978-981-99-1642-9_10

suitable for dealing with linear relationships between signals. Since the development of deep learning, neural networks have been widely used for speech enhancement [6,7]. In recent years, self-encoder structures [8] have been adopted, and methods based on GAN [9–11], DNN [12,13], RNN [14], CNN [15], U-net [16] have also emerged. These deep learning methods are more often used to solve the problem of non-linear relationships between acoustic signals in denoising.

Among deep learning methods, they can be roughly divided into two aspects according to the signal domain in which they work: time domain and time-frequency domain. Our method is based on the second. Early T-F masking methods aimed only at recovering the magnitude of the target speech. After recognizing the importance of phase information, Williamson pioneered the complex Ideal Ratio Mask (cIRM) [17], and their goal is to fully recover the spectrogram of the complex T-F. In Cartesian coordinates, they observed that structure exists in the real and imaginary parts of cIRM, so they devised a DNN-based method to estimate the real and imaginary parts of cIRM.

Yin [18] found that simply changing the training target to cIRM did not recover the phase information, so the parallel branch network used to predict amplitude mask and phase respectively was proposed. Because the phase spectrum in polar coordinates has no structure, Yin added the process of amplitude and phase information exchange in the parallel branch architecture, so that the phase prediction can be guided by the predicted amplitude. At the same time, frequency translation block (FTB) is used to capture the global correlation along the frequency axis, and the harmonics are fully utilized in the DNN model.

However, the speech processed by DNN still has some residual noise that cannot be distinguished by human ears, but some of the methods mentioned previously pay little attention to this point. At the same time, data post-processing methods are usually used to remove residual noise. Therefore, a speech enhancement method named PHASEN-SS was proposed in this paper, which first predicts the magnitude mask and phase, and then post-processes the noisy speech.

This paper achieves three contributions: the first is to combine the time-frequency domain signal processing method with the traditional method; the second is to use the traditional method to post-process the speech data; and the last is to synthesize a new dataset according to different signal-to-noise ratios and noises, and the model is trained and tested on it. A more suitable application scenario is obtained in the synthetic dataset, which reflects the robustness of the model.

2 Related Work

There are three main ideas of this paper: masking methods in the time-frequency domain, phase prediction and spectral subtraction for doing post-processing. This section focuses on the study and application of these three methods.

2.1 Time-Frequency Masking

The key issues to be solved in mask method are mask type and mask prediction. Early T-F masking methods used ideal binary masks (IBM) [19], ideal ratio mask (IRM) [20] or spectral amplitude mask (SMM) [21] to obtain the enhanced amplitude, and finally the enhanced speech was obtained by transformation. Later studies by Paliwal [22] showed that phase plays an important role in speech quality and intelligibility. In order to recover phase, a Phase Sensitive Mask (PSM) was proposed [23]. In addition, the previously mentioned cIRM is a complex-valued mask that can better recover amplitude and phase. But when Williamson uses cIRM, the real and imaginary parts of the cIRM estimated to enhance the amplitude and phase spectra were not as effective as the PSM.

Recently, some methods based on the characteristics of masks have been studied. The supervised DNN estimation ideal ratio mask method was used by Selvaraj [24] in the research of target speech signal enhancement, which designed a SWEMD-VVMDH-DNN model in the network to learn the features of the speech signal, thus reconstructing a noise-free speech signal. Complementary features of multiple masks are used by Zhou [25] to improve speech performance, and the main module of the network is to estimate two complementary templates simultaneously for multi-objective learning. On the other hand, the multi branch extended convolutional network is applied by Zhang [26], and the multi-objective learning framework of complex spectrum and ideal ratio mask is used to enhance the amplitude and phase of speech.

However, the focus of these methods is mainly on the improvement of masking methods, while less attention has been paid to phase features and the connection between phase and mask.

2.2 Phase Prediction

In recent years, phase information has also been deeply studied in speech direction. For example, a complex masking method in polar coordinates is proposed by Choi [27], and the U-Net network with complex depth was used to reflect the distribution of complex ideal ratio masking, and the weighted source distortion rate (wSDR) loss was used to enhance the perception of phase information. A single channel speech enhancement technology based on phase sensitive mask was proposed by Sidheswar [28], which is named PSMGAN. This technique introduces PSM in the end-to-end GAN model and gives importance to the problem of ignoring phase information in traditional end-to-end models. In these methods, they have paid some attention to the phase, but also only studied it in the time domain.

In subsequent research, there are other methods to use amplitude mask and phase information [29–31], and they have also dealt with phase reconstruction asynchronously using amplitude estimation, with the aim of reconstructing the phase based on a given amplitude spectrogram. All these methods show the advantages of phase reconstruction, but they do not make full use of the information in the phase of input noisy in their approaches. So methods for amplitude and phase communication become necessary, and for the implementation of communication methods, the FTB proposed by Yin in the paper plays a key role.

2.3 Spectral Subtraction Processing

In the spectral subtractio [1] usage scenario, the noise is smooth and is additive. It defaults that the first few frames in the noisy signal are ambient noise, and the average amplitude spectrum (energy spectrum) of the first few frames of the noisy signal is taken as the amplitude spectrum (energy spectrum) of the estimated noise. Finally, the amplitude spectrum of the clean signal is obtained by subtracting the estimated noise signal amplitude spectrum from the amplitude spectrum of the noisy signal.

In the process of spectral subtraction, it will judge whether each frame contains speech, then use different methods to process frames containing speech and frames not containing speech. In the absence of speech, the noise spectrum is smoothed and updated to obtain the maximum residual noise value. In the presence of speech, noise cancellation is performed to reduce the residual noise value. This mode of processing with spectral reduction is a better approach to residual noise in the processing of network models and is more appropriate as post-processing of data. In our method, both DNN and spectral subtraction methods are used to better deal with noisy speech.

3 Methods

3.1 Overall Network Structure

The basic idea of PHASEN-SS is that phase and mask features are predicted by two parallel branches respectively. Branch A is used for magnitude mask prediction and branch P is used for phase prediction [18]. The branches are merged at the end of the element-wise product, and finally the residual noise is further removed by spectral subtraction. The overall network structure of PHASEN-SS is shown in Fig. 1.

The spectrogram of noisy speech is used as the input of PHASEN-SS. First, the noisy speech signal is converted from time domain to frequency domain through short-time Fourier transform. The input is denoted as $S^{in} \in R^{T \times F \times 2}$, where T denotes the time step and F denotes the number of frequency bands. Then the graph is input into branch A and branch P respectively. Two different features are generated through different 2D convolutions. The upper part is the amplitude mask prediction, and the obtained features are denoted as $S^A \in R^{T \times F \times CA}$. The second half is the phase prediction, and the obtained features are denoted as $S^P \in R^{T \times F \times CP}$, where CA and CP are the number of channels in branch A and branch P respectively.

The communication process between branch A and branch P is mainly reflected in the component TSB, and the number is set to 3. The exchange of feature information is carried out at the end of each component. The FTB

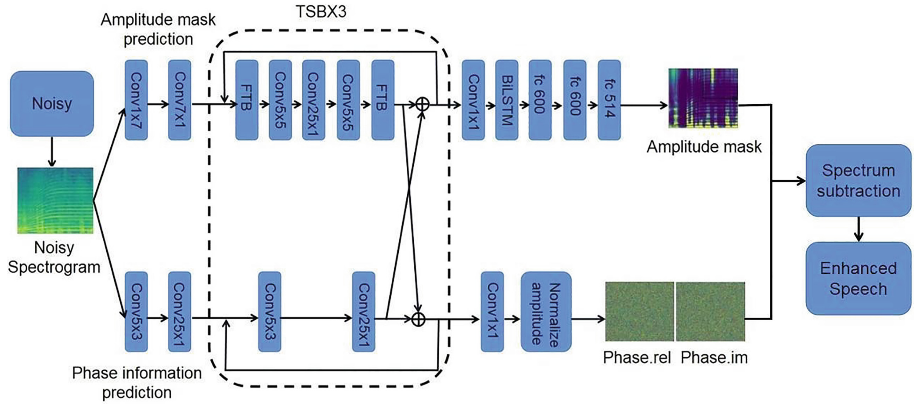


Fig. 1. The overall network structure of PHASEN-SS. The left is the amplitude prediction and the phase prediction, the position shown by the middle dashed box is the communication module (TSB), and the right is the post-processing process.

[18] module is used to extract harmonics in branch A, which is used to capture global correlations along the frequency axis.

At the end of the three TSB modules, for the output of branch A, the channel is reduced to $C_r = 8$ by 1×1 convolution, then reshaped into a 1D feature map, whose dimension is $T \times (F \bullet C_r)$. Finally the feature map is feed Bi-LSTM and three fully connected (FC) to predict an amplitude mask $M \in R^{T\times F\times 1}$. The Sigmaid function is used as the activation function for the last fully connected layer, and ReLU is used as the activation function for the other two fully connected layers. The output of branch P reduces the number of channels to 2 by 1×1 convolution to form a complex value feature map $S^P \in R^{T\times F\times 2}$, and the two channels correspond to the real and imaginary parts of the phase, respectively. Amplitude of this complex feature map is normalized to 1 for each T-F unit, so the feature map only contains phase information. The phase prediction result is denoted by Ψ. Finally, the predicted spectrogram can be computed by the following Eq. (1).

$$S^{out} = abs(S^{in}) \circ M \circ \Psi \tag{1}$$

where $\circ$ represents element-wise multiplication.

3.2 Branch Communication

Branch A: Three 2D convolutional layers are used to handle the local time-frequency correlation of the input features. To obtain the global correlation of the frequency axis, the frequency transform block (FTB) is used before and after the three convolution layers. The combination of 2D convolution and FTB effectively captures global and local correlations, allowing the following blocks to extract high level features for amplitude prediction. The calculation process of branch A is shown in Eq. (2), (3) and (4).

$$S^{A1} = FTB^{in}(S^{A}) \tag{2}$$

$$S^{A2} = conv(S^{A1}), (conv = 3) \tag{3}$$

$$S^{Aout1} = FTB^{out}(S^{A2}) \tag{4}$$

where S^{A} represents the input of the first FTB, S^{A1} is the input of conv and is also the output of the first FTB. The first layer uses a 5 × 5 convolution kernel, the second convolutional layer uses a 25 × 1 convolution kernel, and the third layer uses a 5 × 5 convolution kernel. S^{A2} represents the output of the three convolutional layers and is also the input of the second FTB. S^{Aout1} represents the output of the second FTB and also the input of branch A in the second round of TSB module. This is the flow of the first TSB module of branch A, and TSB loops 3 times.

S^{P} is only processed by two 2D convolution layers in branch P. The execution process is shown in Eq. (5), (6).

$$S^{P1} = S^{P} \tag{5}$$

$$S^{P2} = conv(S^{P1}), (conv = 3) \tag{6}$$

S^{P1} represents the input of the first TSB in the P branch, and conv represents the convolutional layer. The first layer uses a 5 × 3 convolution kernel, and the second convolutional layer uses a 25 × 1 convolution kernel to capture long-range temporal correlations. And global layer normalization (GLN) is performed before each convolutional layer. S^{P2} represents the output of the convolutional layer and is also the input of the next TSB module.

3.3 Spectral Subtraction Denoising

The settings for spectral subtraction are as follows. Let the noisy speech be y(n), the clean speech is x(n), and the noisy speech is e(n).

$$y(n) = x(n) + e(n) \tag{7}$$

Among them, y(n), x(n), and e(n) are the representations of speech in the time domain, which are transformed to the frequency domain by Fourier transform, and are represented as $Y(\omega)$, $X(\omega)$, and $E(\omega)$, respectively.

$$Y(\omega) = X(\omega) + E(\omega) \tag{7}$$

$|\widehat{X}(\omega)|$ is the modulo value of the estimated clean speech, representing the magnitude spectrum of the estimated speech, which is calculated by Eq. (8). When $|\widehat{X}(\omega)|$ is less than 0, it is replaced with 0, where $|Y(\omega)|$ and $|E(\omega)|$ represent $Y(\omega)$ and the modulus of $E(\omega)$.

$$|\widehat{X}(\omega)| = |Y(\omega)| - |E(\omega)| \tag{8}$$

$$|\widehat{X}(\omega)| = \begin{cases} 0, |\widehat{X}(\omega)| < 0 \\ |\widehat{X}(\omega)|, other \end{cases} \tag{9}$$

$\widehat{X}(\omega)$ is the spectrum of the estimated speech, which is obtained by combining the amplitude spectrum of the estimated speech and the phase spectrum of the noisy speech $Y(\omega)$, denoted by $e^{j\varphi Y(\omega)}$ the phase spectrum of $Y(\omega)$. The specific calculation is expressed as Eq. (10).

$$\widehat{X}(\omega) = |\widehat{X}(\omega)| e^{j\varphi Y(\omega)} \tag{10}$$

Finally, the estimated clean speech spectrum is transformed into the time domain through an inverse fourier transform, and the result is the enhanced speech after denoising by spectral subtraction. The specific spectral subtraction denoising processing flow is shown in Fig. 2.

$$\widehat{x}(t) = ISTFT(\widehat{X}(\omega)) \tag{11}$$

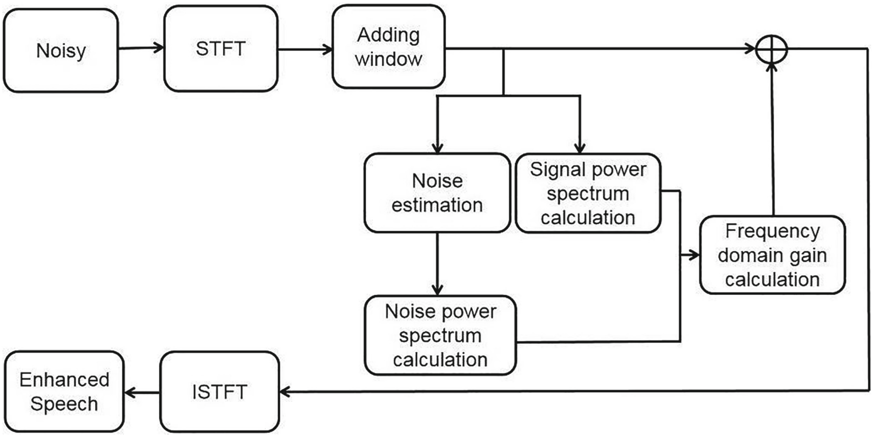

Noisy
STFT
Adding window
Noise estimation
Signal power spectrum calculation
Frequency domain gain calculation
Noise power spectrum calculation
Enhanced Speech
ISTFT

Fig. 2. Spectral subtraction denoising process.

4 Experiments

4.1 Dataset

Two datasets were used in our experiment. The first dataset is Voice Bank + DEMAND [35]: A total of 30 speakers, 28 speakers were included in the training set and 2 speakers were included in the validation set. The training and test sets contain 11,572 and 824 speech pairs, respectively.

The second noisy speech dataset(VB-Noise92) is synthesized with 4 common noises in Noise92 (babble, buccaneer1, factory1, white) and Voice Bank clean speech according to the specified signal-to-noise ratio (–5, 10, 20). The training and test sets are the same size as the first dataset.

4.2 Evaluation Indicators

This paper will use the following four common indicators to evaluate PHASEN-SS and other network models. The higher the score the better.

PESQ: Perceptual assessment of speech quality (from –0.5 to 4.5).

CSIG: Mean Opinion Score (MOS) prediction only involves signal distortion (from 1 to 5) of speech signals.

CBAK: MOS prediction of background noise intrusiveness (from 1 to 5).

COVL: MOS's prediction of the overall effect (from 1 to 5).

4.3 Comparative Experiment

Hardware: Server with graphics card memory size 2080 TI. Software: Linux system, Pytorch platform. Other data preparation: All audio was resampled to 16 kHz, and STFT was calculated using a Hanning window with a window length of 25 ms, a jump length of 10 ms and FFT size of 512. Duing to equipment limitations, both the original model and the improved network model were trained at 6 epochs, with each epoch of size 5786. Adam optimizer with a fixed learning rate of 0.0005 andstep size of 6000 is used, and the batch size is set to 2.

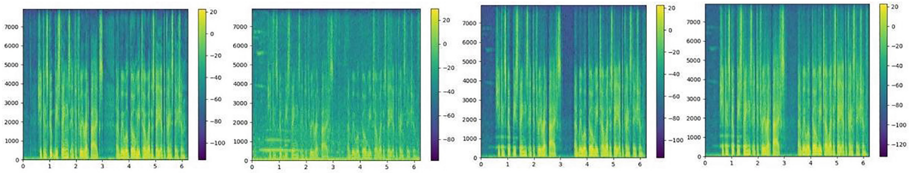
Noisy speech is the baseline of the experiment. Our method is compared with those of the traditional Wiener method and the neural network models SEGAN [9], SASEGAN [32], MMSE-GAN [33], MDPhD [34], PHASEN [18].

The first dataset. Firstly, we compared with Wiener filtering and find that the objective index after Wiener filtering is relatively low, and the neural network model is better, and the effect of PHASE-SS is more obvious. Then we compared with the time-domain methods: SEGAN, SASEGAN and MMSE-GAN. It is found that the T-F domain phase acquisition method is better than the time-domain method in the same data set, and it also proves that the network used in capturing phase-related information is contributing to the results. Finally, we compared with the other two hybrid time-domain and time-frequency domain models (MDPhD, PHASEN), and the objective metrics in the PHASEN results are also slightly higher than these two models, which also indicates that our model is improved to some extent. This shows that spectral subtraction is effective in residual noise removal. All comparison results are shown in Table 1, and PHASEN-SS is our model.

Table 1. Comparison results of different models on Voice Bank +DEMAND

Metric	CSIG	CBAK	COVL	PESQ
Noisy	3.35	2.44	2.63	1.97
Wiener	3.23	2.68	2.67	2.22
SEGAN	3.48	2.94	2.80	2.16
SASEGAN	3.54	3.08	2.93	2.36
MMSE-GAN	3.80	3.12	3.14	2.53
MDPhD	3.85	**3.39**	3.27	2.70
PHASEN	4.02	3.14	3.43	2.83
PHASEN-SS	**4.12**	3.19	**3.52**	**2.91**

The effect of the improved model can also be reflected from the spectrograms of different voices. Figure 3 is a spectrum comparison diagram of each speech: (a) is a clean spectrogram, and (b) is a noisy spectrogram, and (c) is a PHASEN enhanced spectrogram, and (d) is a PHASEN-SS enhanced spectrogram. The spectrogram of all speech can also be clearly seen from the figure, and the spectrogram obtained by the improved model is closer to the spectrogram of clean speech. In terms of details, our results are closer to clean speech than that of PHASEN, while this also shows that the our model has a better effect on speech and the degree of restoration is higher.



(a) Clean (b) Noisy (c) PHASEN (d) PHASEN-SS

Fig. 3. Spectrum comparison chart.

The second dataset. Firstly, when the signal-to-noise ratio is -5, the comparison results enhanced with PHASEN-SS are shown in Table 2. The results of four indicators show that our model has the greatest impact on the noise of buccaneer1, but the effect of comparing the noise-bearing voices with multi-channel mixing (shown underlined) decreases instead. This indicates that the our model is not ideal for noise denoising at low signal-to-noise ratios. It can also be found

Table 2. SNR = –5, 4 kinds of noise comparison results on VB-Noise92

Metric	CSIG	CBAK	COVL	PESQ
DEMAND	<u>4.12</u>	<u>3.19</u>	<u>3.52</u>	<u>2.91</u>
Babble	2.85	2.26	2.16	1.59
Buccaneer1	**2.93**	**2.59**	**2.33**	**1.81**
Factory1	2.63	2.21	2.01	1.50
White	2.72	2.54	2.17	1.71

that the improved model works best for noisy buccaneer1 and worst for noisy factory1 under low SNR.

Then, when the signal-to-noise ratio is 10, the comparison results of using PHASEN-SS enhancement are shown in the Table 3. It can be seen that the indicators CSIG and COVL of the PHASEN-SS method for processing speech with babble noise are improved, and the results of CBAK and PESQ for buccaneer1 are better than those for multi-channel noise speech. The experimental results show that our method improves the CSIG and COVL indexes of bubble and CBAK and PESQ indexes of buccaneer1 respectively when the SNR is medium.

Table 3. SNR = 10, 4 kinds of noise comparison results on VB-Noise92

Metric	CSIG	CBAK	COVL	PESQ
DEMAND	<u>4.12</u>	<u>3.19</u>	<u>3.52</u>	<u>2.91</u>
Babble	**4.19**	3.31	**3.56**	2.93
Buccaneer1	4.03	**3.44**	3.49	**2.95**
Factory1	3.86	3.13	3.25	2.67
White	3.52	3.29	3.10	2.68

Finally, experiments were carried out with a signal-to-noise ratio of 20, and the comparison results using PHASEN-SS enhancement are shown in Table 4. It is also clear from the metrics that the improved model has the most significant enhancement effect on babble-containing data in particular, and its improvement is much better than other speech. This also indicates that our model is more friendly to babble-containing data at high SNR.

Table 4. SNR = 20, 4 kinds of noise comparison results on VB-Noise92

Metric	CSIG	CBAK	COVL	PESQ
DEMAND	4.12	3.19	3.52	2.91
Babble	**4.90**	**3.91**	**4.33**	**3.69**
Buccaneer1	4.63	3.88	4.12	3.57
Factory1	4.62	3.68	4.07	3.51
White	4.07	3.79	3.70	3.30

5 Conclusion

We have utilized two stages to enhance speech: A two-branch network for speech enhancement and spectral subtraction for data post-processing. In this paper, traditional methods are combined with time-frequency domain signal processing methods, and then the speech data are post-processed with traditional methods, and a new noisy dataset is synthesized, on which it is trained and tested. However, there are still some shortcomings in our model, such as no more refined improvement on the DNN model. In the future, we plan to add component loss to the mask prediction of PHASEN-SS to improve the accuracy of mask prediction, and set weighted source distortion rate loss in phase prediction to enhance phase prediction. Finally, we wish our model can be applied to speech recognition.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (No. 61170093).

References

1. Berouti, M, Schwartz, R., Makhoul, J.: Enhancement of speech corrupted by acoustic noise. In: ICASSP IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, pp. 208–211 (1979)
2. Lim, J., Oppenheim, A.: All-pole modeling of degraded speech. IEEE Trans. Acoust. Speech Signal Process. **26**(3), 197–210 (1978)
3. Ephraim Y.: Statistical-model-based speech enhancement systems. In: Proceedings of the IEEE, vol. 80, no. 10, pp. 1526–1555 (1992)
4. Dendrinos, M., Ba Kamidis, S.G., Carayannis, G.: Speech enhancement from noise: a regenerative approach. Speech Commun. **10**(1), 45–57 (1991)
5. Ephraim, Y., Trees, H.V.: A signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process. **3**(4), 251–266 (1995)
6. Tamura, S., Waibel, A.: Noise reduction using connectionist models. In: ICASSP, pp. 553–556 (1988)
7. Parveen, S., Green, P.: Speech enhancement with missing data techniques using recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), pp. 733–736 (2004)
8. Lu, X.G., Tsao, Y., Matsuda, S., et al.: Speech enhancement based on deep denoising autoencoder. In: Conference of the International Speech Communication Association, ISCA, pp. 436–440 (2013)

9. Pascual, S., Bonafonte, A., Serrà, J.: SEGAN: speech enhancement generative adversarial network. Interspeech, 3642–3646 (2017)
10. Abdulatif, S., Armanious, K., Guirguis, K., et al.: Aegan: time-frequency speech denoising via generative adversarial networks. EUSIPCO, pp. 451–455 (2020)
11. Pan, Q., Gao, T., Zhou, J., et al.: CycleGAN with dual adversarial loss for bone-conducted speech enhancement. CoRR.2021:2111.01430
12. Yasuda, M., Koizumi, Y., Mazzon, L., et al.: DOA estimation by DNN-based denoising and dereverberation from sound intensity vector. CORR.2019:1910.04415
13. Yasuda, M., Koizumi, Y., Saito, S., et al.: Sound event localization based on sound intensity vector refined by DNN-based denoising and source separation. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 651–655 (2020)
14. Le, X., Chen, H., Chen, K., et al.: DPCRN: dual-path convolution recurrent network for single channel speech enhancement. In: Interspeech, pp. 2811–2815 (2021)
15. Pandey, A., Wang, D.: 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)
16. Jansson, A., Sackfield, A.W., Sung, C.C.: Singing voice separation with deep u-net convolutional networks: US20210256994A1 (2021)
17. Williamson, D.S., Wang, Y., Wang, D.L.: Complex ratio masking for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(3), 483–492 (2016)
18. Yin, D., Luo, C., Xiong, Z., et al.: Phasen: a phase-and-harmonics-aware speech enhancement network. In: Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence (AAAI).2020: 9458–9465
19. Hu, G., Wang, D.L.: Speech segregation based on pitch tracking and amplitude modulation. In: IEEE Workshop on Applications of Signal Processing to Audio & Acoustics, pp. 553–556 (2002)
20. Srinivasan, S., Roman, N., Wang, D.L.: Binary and ratio time-frequency masks for robust speech recognition. Speech Commun. **48**(11), 1486–1501 (2006)
21. Wang, Y., Narayanan, A., Wang, D.L.: On training targets for supervised speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(12), 1849–1858 (2014)
22. Paliwal, K., Wójcicki, K., Shannon, B.J.: The importance of phase in speech enhancement. Speech Commun. **53**(4), 465–494 (2011)
23. Erdogan, H., Hershey, J.R., Watanabe, S., et al.: Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 708–712 (2015)
24. Selvaraj, P., Eswaran, C.: Ideal ratio mask estimation using supervised DNN approach for target speech signal enhancement. **42**(3), 1869–1883 (2021)
25. Zhou, L., Jiang, W., Xu, J., et al.: Masks fusion with multi-target learning for speech enhancement. Electr. Eng. Syst. Sci. arXiv e-prints (2021)
26. Zhang, L., Wang, M., Zhang, Z., et al.: Deep interaction between masking and mapping targets for single-channel speech enhancement. CORR.2021:2106.04878
27. Choi, H.S., Kim, J.H., Huh, J., et al.: Phase-aware speech enhancement with deep complex U-Net In: ICLR. 2019:1903.03107
28. Routray, S., Mao, Q.: Phase sensitive masking-based single channel speech enhancement using conditional generative adversarial network. Comput. Speech Lang. **71**, 101270 (2021)

29. Takahashi, N., Agrawal, P., Goswami, N., et al.: PhaseNet: discretized phase modeling with deep neural networks for audio source separation. In: Interspeech, pp. 2713–2717 (2018)
30. Takamichi, S., Saito, Y., Takamune, N., et al.: Phase reconstruction from amplitude spectrograms based on von-Mises-distribution deep neural network. In: IEEE 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 286–290 (2018)
31. Masuyama, Y., Yatabe, K., Koizumi, Y., et al.: Deep griffin-lim iteration. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 61–65 (2019)
32. Phan, H., Nguyen, H.L., Chen, O.Y., et al.: Self-attention generative adversarial network for speech enhancement. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7103–7107 (2021)
33. Soni, M.H., Shah, N., Patil, H.A.: Time-frequency masking-based speech enhancement using generative adversarial network. In: ICASSP, pp. 5039–5043 (2018)
34. Kim, J.H., Yoo, J., Chun, S., et al.: Multi-domain processing via hybrid denoising networks for speech enhancement. CoRR.2018:1812.08914
35. Valentini-Botinhao, C., Wang, X., Takaki, S., et al.: Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech. In: 9th ISCA Speech Synthesis Workshop, SSW, pp. 146–152 (2016)