INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Masking Estimation with Phase Restoration of Clean Speech for Monaural Speech Enhancement

Xianyun Wang, Changchun Bao

Speech and Audio Signal Processing Laboratory, Faculty of Information Technology, Beijing University of Technology, Beijing, China, 100124

b201402001@emails.bjut.edu.cn, baochch@bjut.edu.cn

Abstract

Deep neural network (DNN) has become a popular means for separating target speech from noisy speech due to its good performance for learning a mapping relationship between the training target and noisy speech. For the DNN-based methods, the time-frequency (T-F) mask commonly used as the training target has a significant impact on the performance of speech restoration. However, the T-F mask generally modifies magnitude spectrum of noisy speech and leaves phase spectrum unchanged in enhancing process. The recent studies have revealed that incorporating phase spectrum information into the T-F mask can effectively improve perceptual quality of the enhanced speech. So, in this paper, we present two T-F masks to simultaneously enhance magnitude and phase of speech spectrum based on non-correlation assumption of real part and imaginary part about speech spectrum, and use them as the training target of the DNN model. Experimental results show that, in comparison with the reference methods, the proposed method can obtain an effective improvement in speech quality for different signal to noise ratio (SNR) conditions.

Index Terms: Phase restoration, DNN, time-frequency mask, Speech enhancement

1. Introduction

Speech enhancement is very useful due to its wide range of applications such as mobile phone/communication disturbed by background noise, hearing aids and noise-robust speech recognition. Speech enhancement aims at suppressing background noise while improving quality and intelligibility of the enhanced speech.

In the last several decades, some effective methods have been proposed, e.g. spectral-subtractive algorithm [1], Wiener filtering [2] and statistical-model-based method [3]. These methods could achieve a good performance for stationary noise. When the non-stationary noise is concerned, their performance degrades quickly [4]. In order to improve the ability to deal with non-stationary noise, some supervised methods considering pre-training information of speech and noise have emerged and worked well in various noise conditions [4, 5]. For example, the method based on hidden Markov model (HMM) model [5, 6], the method based on auto-regressive (AR) model [4, 7, 8], the method based on Gaussian mixture model (GMM) [9-12], the method based on support vector machine (SVM) [13] and the DNN-based method [14-24]. Specially, the supervised DNN-based speech separation/enhancement methods have achieved a great success, because the strong learning capacity of the DNN can effectively model nonlinear interaction between the training target and the acoustic features of noisy speech. For the DNN-based methods, the learning target plays a very important role on the performance of speech restoration and the T-F mask is commonly used for the learning target. Currently, the mask is mainly focused on restoring magnitude spectrum of clean speech and the recovering of phase spectrum of clean speech is considered rarely. The studies in [25, 26] have shown that the perceptual quality of the enhanced speech can be further improved by restoring phase spectrum. Based on this investigation, a so-called complex ideal ratio mask (cIRM) [22, 23] was proposed. The cIRM is divided into real part and imaginary part. Two real values corresponding to real part and imaginary part are fed into the DNN as the training target simultaneously. Since the imaginary part is considered in the cIRM, the mask can simultaneously restore magnitude and phase spectra.

In this paper, we exploit cosh distance measure [27] to generate a new T-F mask with phase restoration for speech enhancement. Different from the cIRM, the proposed mask is a real-valued mask consisting of two sub-masks. For two sub-masks, one only acts on real part of noisy speech, which is used to restore real part of clean speech spectrum, and the other one only acts on imaginary part of noisy speech for restoring imaginary part of clean speech spectrum. Thus, once these two sub-masks are obtained, the real and imaginary parts of clean speech spectrum can be estimated so that the magnitude and phase spectra of noisy speech are enhanced simultaneously.

The rest of this paper is organized as follows. In Section 2, the structure of complex speech spectrum is described. In Section 3, the details of the proposed method are discussed. Experimental results are provided in Section 4, and the conclusions are given in section 5.

2. The structure of complex speech spectrum

When speech signal is transformed into frequency domain by the short-time Fourier transform (STFT), each T-F unit $X(m,k)$ of the complex spectrum can be expressed as follows,

$$\begin{aligned} X(m,k) &= \left|X(m,k)\right| \cdot \exp\left(j \cdot \varphi(m,k)\right) \\ &= X_r(m,k) + j \cdot X_i(m,k) \end{aligned} \tag{1}$$

where $X_r(m,k)$ and $X_i(m,k)$ are real part and imaginary part of speech spectrum at the k^{th} frequency bin for the m^{th} frame. Thus, the magnitude spectrum and phase spectrum can be represented respectively as follows,

$$\left|X(m,k)\right| = \sqrt{X_r(m,k)^2 + X_i(m,k)^2} \tag{2}$$

http://dx.doi.org/10.21437/Interspeech.2019-1141

and

$$\varphi(m,k)=arctan(\frac{X_i(m,k)}{X_r(m,k)}) \tag{3}$$

where arctan(.) is the arctangent function.

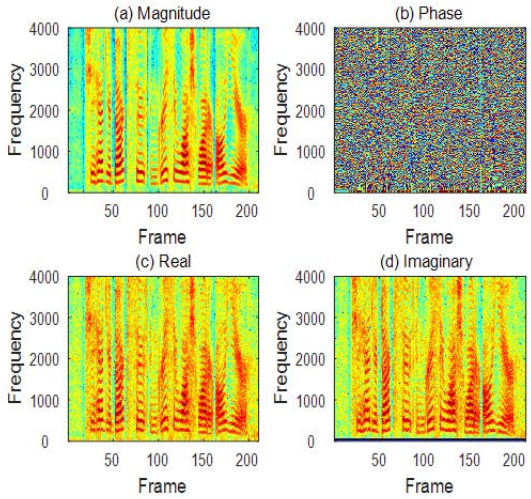

Figure 1: *An example of clean magnitude spectrogram (a), phase spectrogram (b), real spectrogram (c) and imaginary spectrogram (d).*

Fig. 1 gives an example of magnitude spectrogram (top-left), phase spectrogram (top-right), real spectrogram (buttom-left) and imaginary spectrogram (buttom-right) of clean speech signal. From Fig. 1, we can see that the magnitude spectrum, real spectrum and imaginary spectrum all exhibit fine structure and envelope, while the phase spectrum looks like a random spectrum. In order to simultaneously restore the magnitude and phase of speech spectrum, using DNN model to predict magnitude and phase is the most direct way. However, according to the study in [23], the phase of clean speech is hard to be successfully predicted by using DNN model due to the lack of specific structure of spectrum. Due to the similar structure of real and imaginary spectra to the speech magnitude spectrum, naturally, the cIRM was proposed to simultaneously obtain magnitude and phase of speech spectrum [23]. Since the compression of no boundary is possibly detrimental to the DNN model based on gradient descent [18], the cIRM is hard to achieve optimal performance for the unbounded values of real and imaginary components.

3. The proposed approach

In this paper, our goal is to derive a bounded mask applied to the STFT spectrum of noisy speech for producing real and imaginary spectra of clean speech. In other words, given the complex spectrum of noisy speech, $Y(m,k)$, we get complex spectrum estimation of clean speech, $\hat{X}(m,k)$, as follows,

$$\begin{aligned}\hat{X}(m,k) &= H(m,k)\cdot Y(m,k)\\ &= H(m,k)\cdot Y_r(m,k)+j\cdot H(m,k)\cdot Y_i(m,k)\end{aligned} \tag{4}$$

where $H(m,k)$ is the proposed T-F mask. $Y_r(m,k)$ and $Y_i(m,k)$ are real and imaginary parts of noisy speech spectrum, respectively. In the proposed method, the mask $H(m,k)$ is divided two uncorrelated real-valued sub-masks, $H_1(m,k)$ and $H_2(m,k)$, i.e.,

$$H(m,k)=H_1(m,k)+H_2(m,k) \tag{5}$$

For the $H_1(m,k)$ and $H_2(m,k)$, we assume that there is a correlation between the $H_1(m,k)$ and the real part of noisy speech spectrum, and the $H_1(m,k)$ does not affect imaginary part of noisy speech spectrum. Moreover, the $H_2(m,k)$ is only assumed to have an influence on imaginary part of noisy speech spectrum. Note that, different from the cIRM, the $H(m,k)$, $H_1(m,k)$ and $H_2(m,k)$ are all the real-valued masks. Obviously, when the estimation of the $H(m,k)$ is given, the spectral estimation of clean speech, $\hat{X}(m,k)$, can be obtained as follows,

$$\begin{aligned}\hat{X}(m,k) &= H(m,k)\cdot Y(m,k)\\ &= \left(H_1(m,k)+H_2(m,k)\right)\cdot Y_r(m,k)\\ &\quad + j\cdot\left(H_1(m,k)+H_2(m,k)\right)\cdot Y_i(m,k)\\ &= H_1(m,k)\cdot Y_r(m,k)+j\cdot H_2(m,k)\cdot Y_i(m,k)\end{aligned} \tag{6}$$

The study in [2] has pointed out that the Log-MMSE (logarithmic minimum mean square error) measure is more suitable for speech processing. However, the Log-MMSE measure is hard to be used to compare complex spectrum. Fortunately, the author in [27] has given a fact that the performance of the cosh measure is similar to that of the Log-MMSE measure. Thus, in this paper, the cosh measure [27] is used as the cost function to obtain the $H(m,k)$. For the sake of convenience, the index symbols m and k are omitted later. Here, the real and imaginary parts of clean speech spectrum are assumed to be uncorrelated, and the real and imaginary parts of noisy speech spectrum are assumed to be uncorrelated. Moreover, speech and noise are also assumed to be uncorrelated.The cosh measure [27] about H_1 and H_2 is given by

$$\begin{aligned}J(H_1,H_2) &= \frac{\left|X-\hat{X}\right|^2}{X\cdot\hat{X}}=\frac{\left|X-H\cdot Y\right|^2}{X\cdot\left(H\cdot Y\right)}\\ &= \frac{\left|\left(X_r+j\cdot X_i\right)-\left(H_1\cdot Y_r+j\cdot H_2\cdot Y_i\right)\right|^2}{\left(X_r+j\cdot X_i\right)\cdot\left(H_1\cdot Y_r+j\cdot H_2\cdot Y_i\right)}\\ &= \frac{(X_r-H_1\cdot Y_r)^2+(X_i-H_2\cdot Y_i)^2}{(H_1\cdot X_r\cdot Y_r-H_2\cdot X_i\cdot Y_i)}\end{aligned} \tag{7}$$

$J(H_1, H_2)$ is the cost function with respect to two sub-masks, H_1 and H_2. Since the partial derivatives of cost function with respect to the H_1 and H_2 are set to zero, we have

$$H_1=\sqrt{\frac{X_r^2}{X_r^2+N_r^2}} \tag{8}$$

and

$$H_2=\sqrt{\frac{X_i^2}{X_i^2+N_i^2}} \tag{9}$$

where N_r and N_i are real and imaginary parts of noise spectrum, respectively.

Fig.2 gives a comparison about the spectra of the real part (top-left) and imaginary part (top-right) of clean speech, and the real part (bottom-left) and imaginary part (bottom-right) of noisy speech. Herein, noisy speech is generated from babble noise under 0 dB input SNR. Moreover, a comparison about

the spectra of ideal H_1, ideal H_2, the enhanced speech obtained by ideal H_1 and ideal H_2 are shown in Fig.3. From Fig.2 and Fig.3, we can see that using proposed two sub-masks, H_1 and H_2, can effectively remove background noise and restore real part and imaginary part of clean speech spectrum under ideal condition.

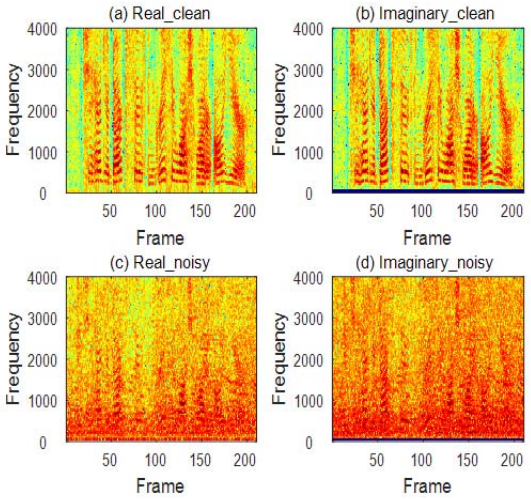

Figure 2: *The spectrogram comparison of clean real part (a), clean imaginary part (b), noisy real part (c) and noisy imaginary part (d).*

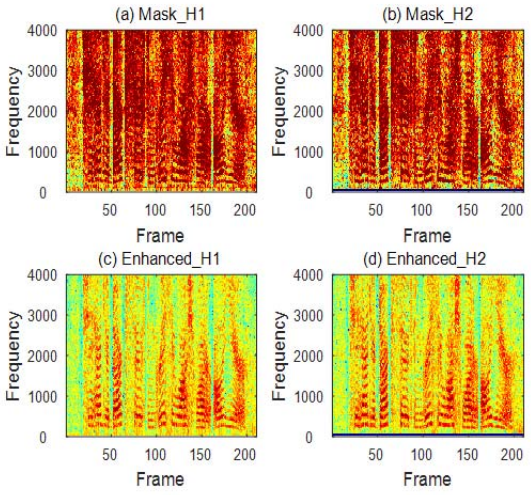

Figure 3: *The spectrogram comparison of mask H_1 (a), mask H_2 (b), enhanced speech obtained by H_1 (c) and H_2 (d).*

Fig. 4 depicts the DNN model that is used to estimate the H_1 and H_2. It is similar to the DNN structure used in [22, 23]. The DNN has three hidden layers and each layer has 1024 nodes with the rectified linear (ReLU) activation function. The backpropagation algorithm with dropout regularization (dropout rate 0.2) and the adaptive gradient descent with a momentum term are used to train the DNN. The momentum rate is 0.5 for the first 5 epochs and 0.9 for the rest epochs. The output layer is separated into two sub-layers, which corresponds to the H_1 and H_2 of the H, respectively. When the proposed mask is used as the learning target, the sigmoid activation function can be used at output layer due to the learning target is in the range of [0, 1].

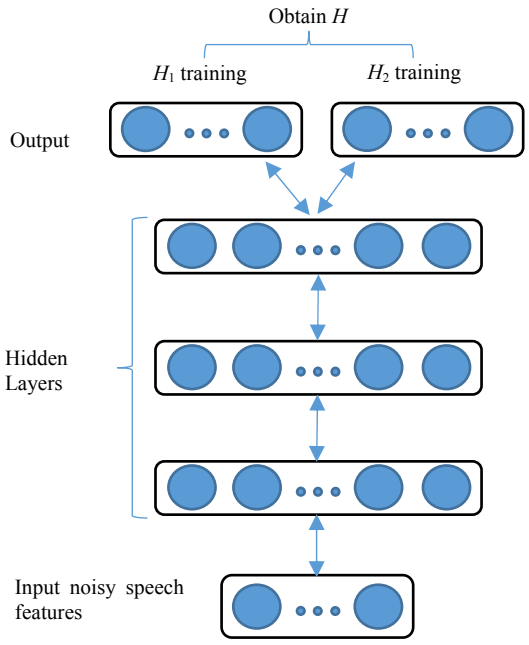

Figure 4: *The basic DNN training diagram.*

4. Experimental results

In the experiment, three different types of background noise are chosen from NOISEX-92 database including babble noise, f16 noise and factory noise for training. In addition, volvo noise from NOISEX-92 database and street noise from ITU-T database are used as the unseen noise for mismatch evaluation. Two hours of utterances are selected from TIMIT training set for the training. In the test, 200 utterances are chosen from the TIMIT test set. The noisy speech is obtained by adding above-mentioned five types of noise to speech signal at four input SNR levels (i.e., -5dB, 0dB, 5dB and 10dB). The speech and noise are down-sampled to 8 kHz. The frame length is 256 samples and the frame shift is 128 samples. In this work, the DNN-based method given in [21] is selected as the first reference method (named Ref.1). The cIRM incorporating phase information [22] is considered as the second reference method (named Ref.2). The proposed masking estimation method is termed as the Pro. In training stage, the DNN model of all methods is composed of three hidden layers, each layer has 1024 hidden units and the log-power spectrum of noisy speech is used as input feature of the DNN model. The mean squared error is used as the cost function for the DNN training. The number of output units correspond to the dimensionality of the training target. In the output layer, the sigmoid activation function is selected in the Ref.1 and proposed method, while linear activation function is selected in the Ref.2. The quality measurements used in this paper include perceptual evaluation of speech quality (PESQ) [28] and short-time objective intelligibility (STOI) [29].

Table 1 shows the PESQ results for five types of noise at different input SNR levels. It is clear that the PESQ of the Ref.1 without speech phase information is relatively lower than the Ref.2 and proposed method in all input SNR conditions. This implies that the masking methods including phase information can recover more spectral information of speech to result in quality improvement. For the Ref.2, the real and imaginary parts of noisy speech spectrum are compressed

during the DNN training, this compressed cIRM may be detrimental to the supervised approaches based on the gradient descent [18] so that the ability of improving PESQ of the Ref.2 is weaker than the proposed method.

Table 1: *Test Results of the PESQ.*

noise type	input SNR	Methods			
		Noisy	Ref.1	Ref.2	Pro
Babble	-5dB	1.719	1.885	1.896	1.971
	0dB	1.913	2.241	2.249	2.259
	5dB	2.220	2.608	2.640	2.649
	10dB	2.518	2.916	2.948	2.981
F16	-5dB	1.595	1.994	2.011	2.131
	0dB	1.832	2.340	2.418	2.485
	5dB	2.136	2.612	2.664	2.789
	10dB	2.444	2.905	2.971	3.073
Factory	-5dB	1.761	2.225	2.292	2.332
	0dB	2.073	2.608	2.645	2.675
	5dB	2.391	2.925	2.977	3.004
	10dB	2.705	3.182	3.214	3.280
Street	-5dB	1.990	2.369	2.384	2.461
	0dB	2.320	2.773	2.799	2.863
	5dB	2.649	3.114	3.169	3.198
	10dB	2.973	3.367	3.433	3.476
Volvo	-5dB	2.818	3.215	3.229	3.283
	0dB	3.121	3.436	3.472	3.512
	5dB	3.446	3.598	3.661	3.682
	10dB	3.759	3.686	3.707	3.799

Fig. 5 shows the average PESQ results of the proposed method and reference methods at different input SNR levels. It is clear that the PESQ of the Ref.1 without speech phase information is relatively lower than the Ref.2 and proposed method in all input SNR conditions. This implies that the masking methods including phase information can recover more spectral information of speech to result in quality improvement. For the Ref.2, the real and imaginary parts of noisy speech spectrum are compressed during the DNN training, this compressed cIRM may be detrimental to the supervised approaches based on the gradient descent [18] so that the ability of improving PESQ of the Ref.2 is weaker than the proposed method.

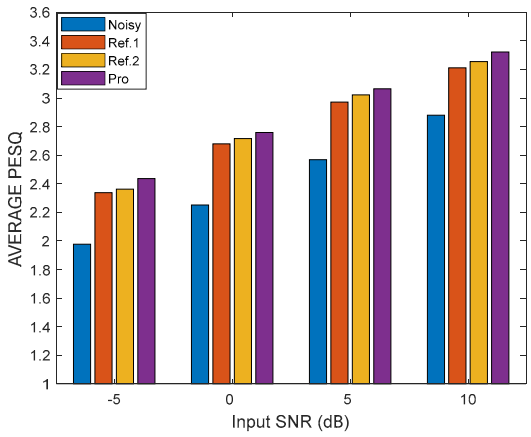

Figure 5: *The average PESQ of difference noises.*

Table 2: *Test Results of the STOI (%).*

Noise type	input SNR	Methods			
		Noisy	Ref.1	Ref.2	Pro
Babble	-5dB	48.02	53.51	53.44	53.69
	0dB	60.40	67.90	68.09	68.68
	5dB	72.24	78.30	79.22	79.51
	10dB	81.92	84.87	85.92	86.47
F16	-5dB	49.46	62.59	63.55	65.89
	0dB	62.32	73.35	75.71	77.33
	5dB	74.55	81.10	83.76	84.79
	10dB	84.26	86.79	88.66	89.60
Factory	-5dB	59.54	70.09	70.06	71.14
	0dB	70.77	79.21	80.21	80.75
	5dB	80.15	85.20	86.45	86.91
	10dB	87.26	89.15	90.73	90.87
Street	-5dB	69.59	76.06	75.99	78.13
	0dB	77.79	82.97	83.23	84.94
	5dB	84.05	87.45	88.35	89.19
	10dB	88.79	90.34	91.21	91.90
Volvo	-5dB	84.81	87.35	87.36	88.94
	0dB	88.45	89.95	91.16	91.40
	5dB	91.55	91.72	92.71	93.15
	10dB	93.05	92.85	94.01	94.32
Average	-5dB	62.28	69.92	70.08	71.55
	0dB	71.94	78.67	79.68	80.62
	5dB	80.50	84.75	86.09	86.71
	10dB	87.25	88.80	90.10	90.63

Table 2 gives the STOI results of different methods at different input SNRs. The Ref.1, Ref.2 and proposed method can effectively improve speech intelligibility compared with noisy speech in most cases. The masking methods considering speech phase, e.g., the Ref.2 and proposed method, are superior to the Ref.1. This proved that incorporating phase information into mask is helpful to increase intelligibility. As a comparison, the proposed system gives a higher STOI value than the Ref.2.

5. Conclusions

In this paper, a T-F masking method was presented by considering both magnitude and phase spectrum information for single-channel speech enhancement. Different from the existing mask with phase information, the proposed method divides the conventional real-valued mask into two uncorrelated real-valued sub-masks. One of them is only assumed to be related to real part of noisy speech spectrum, which is used to restore real part of clean speech spectrum, and the other one is only assumed to be related to imaginary part of noisy speech spectrum for restoring imaginary part of clean speech spectrum. Since the real and imaginary parts of clean speech spectrum are all estimated, the magnitude and phase information can be restored simultaneously. The quality and intelligibility tests showed that the proposed method outperformed reference methods.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61831019, No. 61471014, and No. 61231015).

7. References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 27, no. 2, pp. 113-120, 1979.

[2] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, 670 FL, USA: CRC Press, 2007.

[3] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics Speech and Signal Process*., vol. 32, no. 6, pp. 1109-1121, 1984.

[4] S. Srinivasan, J. Samuelsson, W. B. Kleijin, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 2, pp. 441-452, 2007.

[5] D. Y. Zhao, W. B. Kleijin, "HMM-based gain modeling for enhancement of speech in noise," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 882-892, 2007.

[6] F. Deng, C. C. Bao, W. B. Kleijin, "Sparse hidden Markov models for speech enhancement in non-stationary noise environments," *IEEE Trans. Audio, Speech, Language Process*., vol. 23, no.11, pp. 1973-1987, 2015.

[7] S. Srinivasan, J. Samuelsson, W. B. Kleijin, "Codebook driven short term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no.1, pp. 163-176, 2006.

[8] X. Y. Wang, C. C. Bao, "Speech enhancement using a joint MAP estimation of LP parameters," *Int. conf. on signal process., comm., and comput.*, Ningbo, China, 2015.

[9] A. Reddy, B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1766-1776, 2007.

[10] M. H. Radfar, R. M. Dansereau, "Single-channel speech separation using soft masking filtering," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no.8, pp. 2299-2310, 2007.

[11] K. Hu, D. L. Wang, "An iterative model-based approach to cochannel speech separation," *EURASIP J. Audio, Speech, Music Process.*, vol. 14, pp.1-1, 2013.

[12] G. Kim, Y. Lu, Y. Hu, P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *Journal of the Acoust. Society of Amer.*, vol. 126, pp. 1486-1494, 2009.

[13] K. Han, D. L. Wang, "A classification based approach to speech segregation," *Journal of the Acoust. Society of Amer.*, vol. 132, pp. 3475-3483, 2012.

[14] X. Lu, Y. Tsao, S. Matsuda, C. Hori, "Speech enhancement based on deep denoising autoencoder," *in Proc. Interspeech*, Lyon, France, 2013, pp. 436-440, 2013.

[15] B. Y. Xia, C. C. Bao, "Wiener filtering based speech enhancement with Weighted Denoising Auto-encoder and noise classification," *Speech Comm.*, vol. 60, pp. 13-29, 2014.

[16] Y. Xu, J. Du, L. Dai, C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Lett.*, vol. 21, no.1, pp. 66-68, 2014.

[17] Y. Xu, J. Du, L. Dai, C. Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 23, no.1, pp. 7-19, 2015.

[18] Z. Wang, X. Wang, X. Li, Q. Fu, Y. Yan, "Oracle performance investigation of the ideal masks," *in IWAENC*, Xi'an, China, pp. 1-5, 2016.

[19] J. Chen, Y. Wang, D. L. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 22, pp. 1993-2002, 2014.

[20] M. Delfarah, D. L. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 25, pp. 1085-1094, 2017.

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

[22] D. S. Williamson, Y. X. Wang, D. L. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," *in Proc. ICASSP*, Shanghai, China, pp. 5220-5224, 2016.

[23] D. S. Williamson, Y. X. Wang, D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 24, pp. 483-492, 2016.

[24] D. L. Wang, J. Chen, "Supervised speech separation based on deep learning: an overview," *IEEE Trans. Audio Speech Lang. Process.*, vol. 26, no. 10, pp. 1702-1726, 2018.

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

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

[27] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the speech magnitude spectrum," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857-869, 2005.

[28] 'Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs,' ITU-T Recommendation, P.862, Feb, 2001.

[29] C. H. Taal, R. C. Hendriks, R. Heusdend, et al. "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 19, no. 7, pp. 2125-2136, 2011.