Speech Enhancement with Phase Correction based on Modified DNN Architecture

Rui Cheng, Changchun Bao, Yang Xiang
Speech and Audio Signal Processing Laboratory, Beijing University of Technology, Beijing, China
E-mail: chengrui@emails.bjut.edu.cn, baochch@bjut.edu.cn, xiangyang3131777@emails.bjut.edu.cn

***Abstract*— Speech enhancement is an important issue in the field of speech signal processing. With the development of deep learning, speech enhancement technology combined with neural network has provided a more diverse solution for this field. In this paper, we present a new approach to enhance the noisy speech, which is recorded by a single channel. We propose a phase correction method, which is based on the joint optimization of clean speech and noise by deep neural network (DNN). In this method, the ideal ratio masking (IRM) is employed to estimate the clean speech and noise, and the phase correction is combined to get the final clean speech. Experiments are conducted by using TIMIT corpus combined with four types of noises at three different signal to noise ratio (SNR) levels. The results show that the proposed method has a significant improvement over the referenced DNN-based enhancement method for both objective evaluation criterion and subjective evaluation criterion.**

I. Introduction

In speech signal processing, speech enhancement is mainly used to remove various types of noise in noisy speech. It concerns how to improve speech quality and intelligibility. Over the years, with the development of signal processing technology, various speech enhancement methods have been proposed, such as spectral subtraction [1], Wiener filtering [2], subspace methods [3], statistical model-based methods [4] and so on. These methods have been widely used due to their clear principles and easy implementation. They have become the classic methods in the field of speech enhancement. However, there are two drawbacks to these methods. Firstly, these methods are all based on the estimation of power spectral density (PSD) of the noise, and the accuracy of noise PSD estimation has become the bottleneck of these methods, especially in the case of non-stationary noise or low SNR conditions. Secondly, these methods only enhance the magnitude spectrum and use the phase of noisy speech to reconstruct clean speech, so they ignore the importance of phase information [5][6].

In recent years, with the advance of deep learning, more and more scholars have begun to introduce deep learning into speech signal processing. Especially in speech enhancement, deep learning has achieved state-of-the-art results. Y. Xu et al. proposed a DNN-based speech enhancement method that directly mapped the clean speech from the noisy speech and used the phase of the noisy speech to reconstruct clean speech [10]. Wang et al. proposed a speech enhancement method based on IRM and DNN [7][8][9]. In their methods, DNN was used to learn to map the relationship between the noisy speech and target feature. In addition, the supervisory signal of these methods is the ideal IRM value or the clean speech. By combining the estimated IRM with the phase of the noisy signal, an enhanced speech signal is obtained. Although the deep learning-based approaches achieve better performance than the traditional methods, they still have significant defective in the estimation of the magnitude spectrum and use the phase of the noisy speech as the phase of the enhanced speech, thereby ignoring the importance of the phase information of the speech signal.

With further research, at the same time, some researchers have begun to pay more attention to the role of phase in speech enhancement. Kuldip Pailwal pointed out the importance of phase information in speech enhancement in his work [13] and indicated that the method of phase spectrum compensation could considerably improve the speech quality. Timo Gerkmann proposed a phase reconstruction algorithm based on short-time Fourier transform [24]. Wang et al. proposed a method of DNN-based complex ratio masking, which converted the magnitude and phase into a complex form and used the real and imaginary parts of the speech short-time Fourier transform as the feature signal [23]. Z. Wenlu et al. proposed a modified wiener filtering speech enhancement algorithm with phase spectrum compensation [14]. In that work, the phase compensation method combining with the traditional wiener filtering-based speech enhancement method achieved a good enhancement performance. Therefore, the phase has become an indispensable part of speech enhancement.

In this paper, we propose an IRM-based speech enhancement method that includes the DNN and phase correction. In our approach, we use an IRM-based DNN structure to estimate the magnitude spectrum of clean speech and noise jointly. Then, by applying the noisy speech and the estimated magnitude spectrum of noise, the correction of phase compensation is performed. Therefore, our method not only estimates the magnitude of the clean speech more accurately under the multi-tasking learning, but also produces more accurate phase information for speech reconstruction.

The structure of the paper is organized as follows: Section II describes the proposed enhancement method. Experiments and analysis are shown in Section III. Finally, Section IV summarizes and looks forward to our work.

978-988-14768-5-2 ©2018 APSIPA

II. Proposed Method

A. Proposed Model

In this paper, a new DNN-based architecture for monaural speech enhancement is proposed, which jointly optimize the magnitude spectrum of the clean speech and noise with ideal ratio mask and DNN. The proposed DNN architecture is a mapping of the logarithmic power spectrum (LPS) of noisy speech and the magnitude spectrum of clean speech and noise. The proposed DNN architecture is illustrated in Fig. 1. As shown in Fig. 1, the magnitude spectrum and logarithmic power spectrum of noisy speech are obtained through feature extraction. The LPS is used for the input feature of the proposed DNN. The magnitude spectrum of noisy speech is combined with the estimation of IRM to calculate the magnitude spectrum of the clean speech and noise.

The original output layer of the network has two parts, which are clean speech magnitude spectrum $|\tilde{\mathbf{S}}|$ and noise magnitude spectrum $|\tilde{\mathbf{N}}|$, respectively. We use a pseudo output layer as the output of the network in order to get more accurately enhanced speech. It still contains two outputs and they are enhanced clean speech magnitude spectrum $|\hat{\mathbf{S}}|$ and enhanced noise magnitude spectrum $|\hat{\mathbf{N}}|$. In the pseudo output layer, we use IRM as a transfer function. The typical IRM [12] is usually expressed as follows:

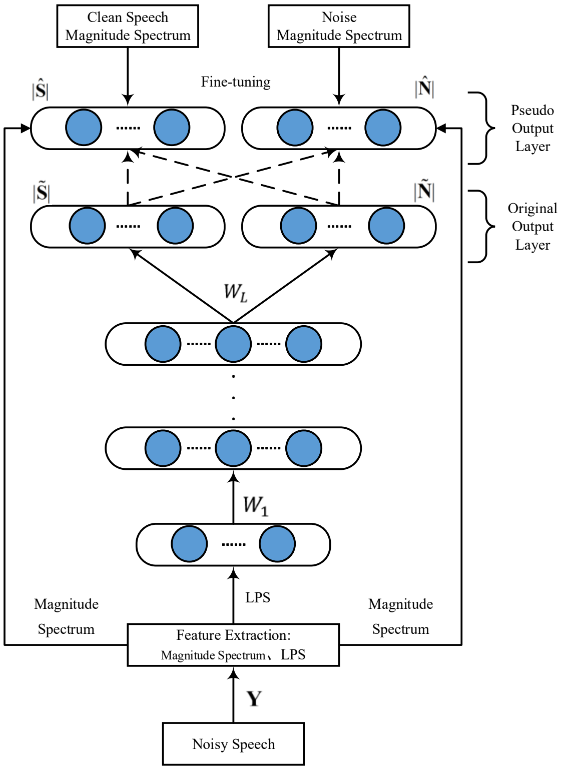

Fig. 1 The architecture of the proposed DNN

$$IRM=\frac{|\mathbf{S}(\omega)|^2}{|\mathbf{S}(\omega)|^2+|\mathbf{N}(\omega)|^2} \tag{1}$$

where $|\mathbf{S}(\omega)|^2$ and $|\mathbf{N}(\omega)|^2$ denote speech energy and noise energy at the frequency index ω, respectively. It is similar to the classical Wiener Filter, which is the optimal estimator of clean speech in the power spectrum.

When the characteristic signal of the noisy speech passes through the proposed DNN, clean speech and noise signals are obtained at the original output layer. In order to estimate the enhanced clean speech and noise more accurately, we use the IRM with constraint as transfer function for the pseudo output layer, and the clean speech transfer function of the pseudo output layer based on [25] is expressed as follows:

$$sIRM=\frac{|\tilde{\mathbf{S}}(\omega)|^2}{|\tilde{\mathbf{S}}(\omega)|^2+\mu|\tilde{\mathbf{N}}(\omega)|^2} \tag{2}$$

where $|\tilde{\mathbf{S}}(\omega)|$ and $|\tilde{\mathbf{N}}(\omega)|$ denote clean speech magnitude spectrum and noise magnitude spectrum from original output layer at the frequency index ω, respectively. And the μ is a constraint factor, it is given by [25]:

$$\mu=\begin{cases}\mu_0\text{-}(SNR_{dB})/s & -5<SNR_{dB}<20\\ 1 & SNR_{dB}\geq 20\\ \mu_{\max} & SNR_{dB}\leq -5\end{cases} \tag{3}$$

where μ_{max} is the upper limit of μ, set to 10, and μ_0=(1+4μ_{max})/5, s=25/(μ_{max}-1). And the expression of SNR_{dB} is as follows:

$$SNR_{dB}=10\log_{10}\frac{\sum|\tilde{\mathbf{S}}(\omega)|^2}{\sum|\tilde{\mathbf{S}}(\omega)|^2} \tag{4}$$

So, we obtain the reconstructed magnitude spectrum of enhanced speech by multiplying the *sIRM* and the magnitude spectrum of noisy speech, it can be expressed as follows:

$$\begin{aligned}|\hat{\mathbf{S}}(\omega)| &= sIRM\odot|\mathbf{Y}(\omega)|\\ &=\frac{|\tilde{\mathbf{S}}(\omega)|^2}{|\tilde{\mathbf{S}}(\omega)|^2+\mu|\tilde{\mathbf{N}}(\omega)|^2}\odot|\mathbf{Y}(\omega)|\end{aligned} \tag{5}$$

where $\odot$ is the element-wise multiplication, and $|\mathbf{Y}(\omega)|$ is the noisy speech magnitude spectrum.

Just like the estimation of clean speech, we also use the same method to estimate the noise signal. And based on the assumption that noise is additive, we can derive as follows:

$$\begin{aligned}|\hat{\mathbf{N}}(\omega)| &= |\mathbf{Y}(\omega)|-|\hat{\mathbf{S}}(\omega)|\\ &=|\mathbf{Y}(\omega)|-\frac{|\tilde{\mathbf{S}}(\omega)|^2}{|\tilde{\mathbf{S}}(\omega)|^2+\mu|\tilde{\mathbf{N}}(\omega)|^2}\odot|\mathbf{Y}(\omega)|\\ &=\frac{\mu|\tilde{\mathbf{N}}(\omega)|^2}{|\tilde{\mathbf{S}}(\omega)|^2+\mu|\tilde{\mathbf{N}}(\omega)|^2}\odot|\mathbf{Y}(\omega)|\end{aligned} \tag{6}$$

Thus, we can calculate the noise ideal ratio mask *nIRM*, which is able to be written as:

$$nIRM = \frac{\mu\left|\tilde{\mathbf{N}}(\omega)\right|^2}{\left|\tilde{\mathbf{S}}(\omega)\right|^2 + \mu\left|\tilde{\mathbf{N}}(\omega)\right|^2} \tag{7}$$

In order to obtain a more accurate estimate of clean speech and noise, we simultaneously supervise the enhanced speech and noise, turning the speech enhancement problem into a multi-task training problem for neural networks. We choose the Mean Squared Error (MSE) as the objective function, and it is minimized between the enhanced speech magnitude spectrum and clean speech magnitude spectrum as well as the enhanced noise magnitude spectrum and real noise magnitude spectrum. So, the objective function is formed as:

$$loss = \left\| \left|\hat{\mathbf{S}}(\omega)\right| - \left|\mathbf{S}(\omega)\right| \right\|_2^2 + \left\| \left|\hat{\mathbf{N}}(\omega)\right| - \left|\mathbf{N}(\omega)\right| \right\|_2^2 \tag{8}$$

where $\|\cdot\|_2$ denotes L2 norm.

B. Phase Correction

The general speech enhancement method uses the phase of the noisy speech as the phase of the enhanced speech. In this way, the resulting error can be ignored when the SNR is high. However, if the SNR is at a relatively low level, the introduced error will cause a certain amount of negative impact for the enhanced speech. Therefore, in this subsection, we introduce the phase correction method and combine it with the proposed network to jointly improve the quality of enhanced speech.

According to the characteristics that the noisy speech signal is a real-valued signal, it is easy to deduce that its short-time Fourier transform is conjugate symmetric. To compensate for the phase of noisy speech, we use a time and frequency dependent phase spectrum compensation function $\mathbf{\Lambda}_k(\omega)$ to adjust noisy speech's STFT value. It is given by [14]:

$$\mathbf{\Lambda}_k(\omega) = \beta T_k \left|\hat{\mathbf{N}}_k(\omega)\right| \tag{9}$$

where β is the compensation factor, and $|\hat{\mathbf{N}}_k(\omega)|$ is the enhanced magnitude spectrum of the noise at the kth frame, T_k is the time-invariant anti-symmetry function and it is given by [14]:

$$T_k = \begin{cases} 1 & 0 < {}^{k}\!/_{N} < 0.5 \\ -1 & 0.5 < {}^{k}\!/_{N} < 1 \\ 0 & else \end{cases} \tag{10}$$

where N is the length of the window in the STFT.

In order to adapt the compensation function to various noises better, the proposed method defines the compensation factor β as a function of the input SNR of noisy speech, and its expression is given as follows [14]:

$$\beta = ce^{-\frac{|\mathbf{Y}_k(\omega)|^2}{|\mathbf{N}_k(\omega)|^2}} \tag{11}$$

where c is a constant, and $|\mathbf{Y}_k(\omega)|$ is the short-time magnitude spectrum of the noisy at the kth frame. Finally, we can get the phase compensation function as follow:

$$\mathbf{\Lambda}_k(\omega) = ce^{-\frac{|\mathbf{Y}_k(\omega)|^2}{|\mathbf{N}_k(\omega)|^2}} T_k \left|\hat{\mathbf{N}}_k(\omega)\right| \tag{12}$$

Since the magnitude spectrum of noise is symmetric, after being corrected by (12), we obtain an anti-symmetric function $\mathbf{\Lambda}_k(\omega)$. This anti-symmetric function $\mathbf{\Lambda}_k(\omega)$ is used to cancel the short-time spectrum of noisy speech as follow [14]:

$$\mathbf{Y}_{\mathbf{\Lambda}_k}(\omega) = \mathbf{Y}_k(\omega) + \mathbf{\Lambda}_k(\omega) \tag{13}$$

The corrected phase spectrum can be derived through:

$$\theta_{\mathbf{Y}_{\mathbf{\Lambda}_k}} = \angle \mathbf{Y}_{\mathbf{\Lambda}_k}(\omega) \tag{14}$$

Therefore, after the corrected phase is combined with the enhanced magnitude spectrum of speech from the network, the corrected enhanced speech spectrum can be written by

$$\hat{\mathbf{S}}_k(\omega) = \left|\hat{\mathbf{S}}_k(\omega)\right| \cdot e^{j\theta_{\mathbf{Y}_{\mathbf{\Lambda}_k}}} \tag{15}$$

C. Proposed Speech Enhancement System

The diagram of the proposed method is shown in Fig. 2. Like the general DNN-based supervised problem, the DNN-based speech enhancement problem also includes a training stage and an enhancement stage.

In the training stage, we map relationship between LPS of noisy speech and magnitude spectrum of clean speech and noise to train DNN with ideal ratio mask. In the enhancement stage, the LPS of the noisy speech passes through the trained DNN to obtain an enhanced magnitude spectrum of speech and noise. The estimated magnitude spectrum of noise and noisy speech are used to further correct the phase of the noisy speech and the corrected phase is combined to perform an inverse Fourier transform for obtaining an enhanced speech.

III. Experiments

A. Preparation of the Dataset

In the experiments, we use the TIMIT [16] corpus to evaluate the speech enhancement performance of proposed method. We use 4620 sentences from different speakers in the TIMIT corpus as the clean speech of training set. The 102 noise types included 100 environmental noise [17], *Babble* and *F16* noise [18], which are used as the noise of training set. All of the signals are down-sampled to 8kHz. Moreover, the 4620 sentences of clean speech and noise of 102 types are artificially mixed at four different SNR levels from -5 to 10dB spaced by 5dB. We randomly selected 9216 sentences of the mixed speech to build an 8-hour training set of noisy speech.

In the enhancement stage, another 201 sentences from the TIMIT test set are randomly chosen as the clean speech of testing set. The two noises (*Babble* and *F16*) in the training set and other two noises (*Factory* and *Street* [18]) outside the training set are combined with the clean speech to get noisy speech for testing. The noisy speech is formed under three different SNR levels ranging from 0 to 10dB at step of 5dB. Additionally, the length of the testing set is about 10 minutes.

B. Network Parameter Settings

The proposed speech enhancement method incorporating phase correction and modified DNN architecture is used to make a comprehensive comparison with several classic state-of-the-art DNN-based speech enhancement methods, they are the baseline DNN (B-DNN) method and IRM-based DNN (IRM-DNN) method:

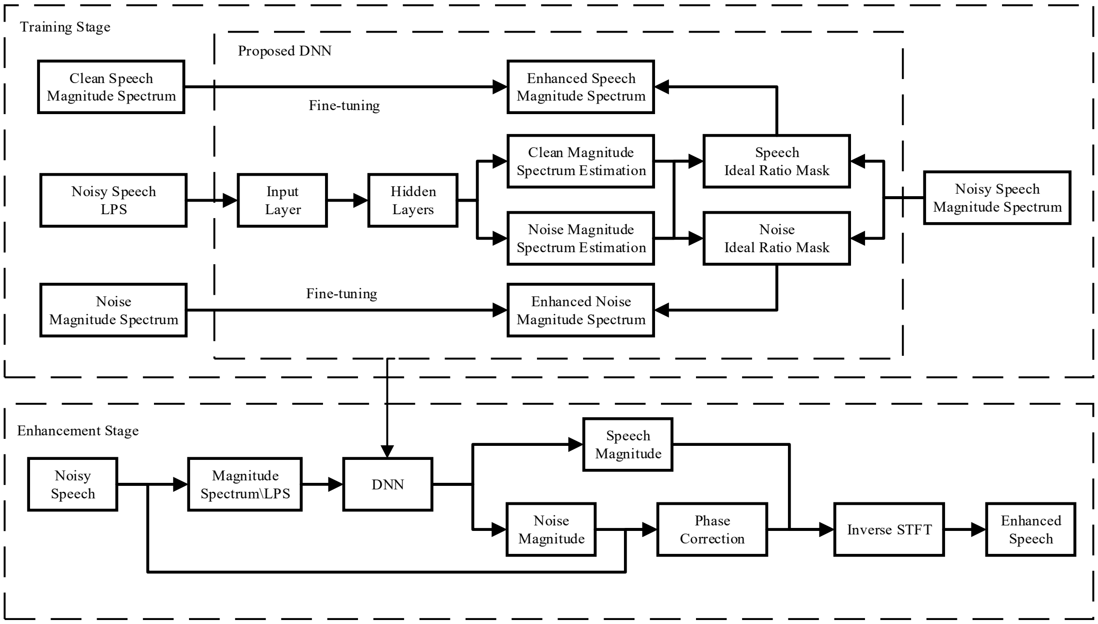

Fig. 2 The diagram of the proposed speech enhancement system

- B-DNN: A DNN-based speech enhancement method that directly maps the LPS of noisy speech signal to the magnitude spectrum of clean speech signal, and use the phase of the noisy speech signal [10].
- IRM-DNN: A DNN-based speech enhancement method that uses an IRM as a constraint to map the LPS of noisy speech signal through a DNN to the magnitude spectrum of clean speech, and use the phase of the noisy speech signal [15].

In our experiments, the structure of all the models is composed of three hidden layers and each layer contains 2048 neurons. We use 129-dimensional speech power spectrum as the input features, which are extracted using a window length of 32ms (256 samples) and a frame shift of 16ms (128 samples). For each neuron in the network, we use Rectified Linear Unit (ReLU) as an activation function to avoid the vanishing gradient problem when the network training using random initialization. The Adaptive Moment Estimation (Adam) algorithm [11] is chosen to update the parameters of the neural network.

C. Evaluation Metrics

The performance of enhanced speech is evaluated by perceptual evaluation of speech quality (PESQ) [19], the extended short-time objective intelligibility (ESTOI) [20] and the segment SNR (SSNR) [21]. The PESQ is used to measure the subjective quality of speech. The higher its value, the better the subjective quality of speech is. The ESTOI measures speech intelligibility. In contrast to STOI [22], extended STOI (ESTOI) does not assume mutual independence between frequency bands. Moreover, ESTOI also incorporates spectral correlation by comparing complete 400-ms length spectrograms of the noisy/enhanced speech and the clean speech signals. As a result, ESTOI can better reflect the intelligibility of speech. The SSNR is an objective measure of speech quality. It is defined by [21]

$$SSNR = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Nm}^{Nm+N-1} x^2(n)}{\sum_{n=Nm}^{Nm+N-1}(x(n)-\hat{x}(n))^2} \tag{16}$$

where *x(n)* and *x̂(n)* denote clean speech and enhanced speech at the time index *n*, respectively, *N* is the frame length (usually 15-20ms is selected) and *M* is the number of frames used for test.

D. Experimental Results and Discussions

Table I presents the average PESQ scores, Fig. 3 shows ESTOI scores and Fig. 4 shows the SSNR results of the proposed DNN method with phase correction (we call it PC-DNN), B-DNN method and the IRM-DNN method at three different SNR levels (0dB, 5dB, 10dB). For the four noises (*Babble*, *F16*, *Factory*, *Street*) test, the *Babble* and *F16* noises are included in the training set, while the *Factory* and *Street* noises, which are common in life, are not included in the training set.

TABLE I. THE AVERAGE PESQ SCORES

SNR (dB)	**Methods**			
	Noisy	***B-DNN***	***IRM-DNN***	***PC-DNN***
0	2.0394	2.3990	2.4465	**2.4872**
5	2.3496	2.6500	2.8058	**2.8499**
10	2.6602	2.8541	3.1328	**3.1709**

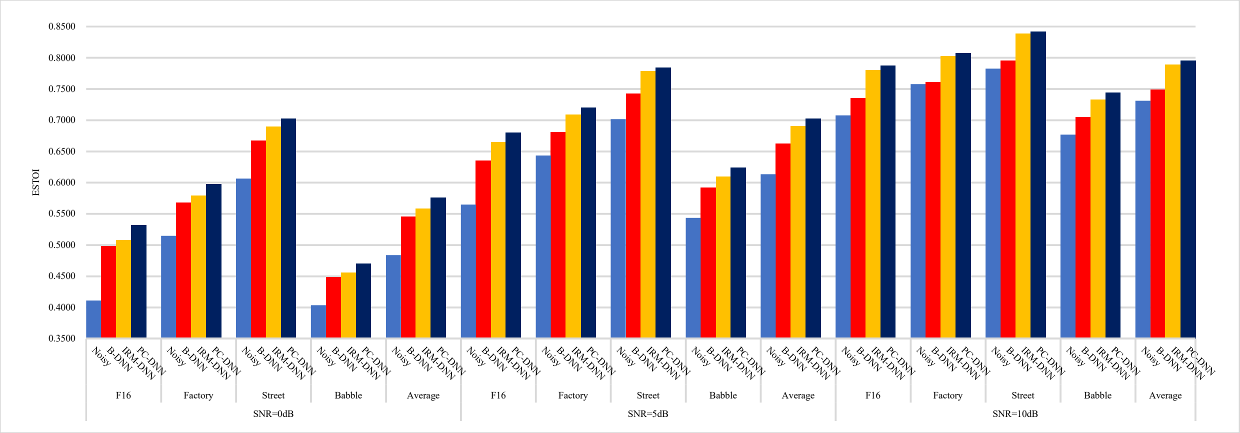

Fig. 3 ESTOI comparison for four noises under three input SNRs

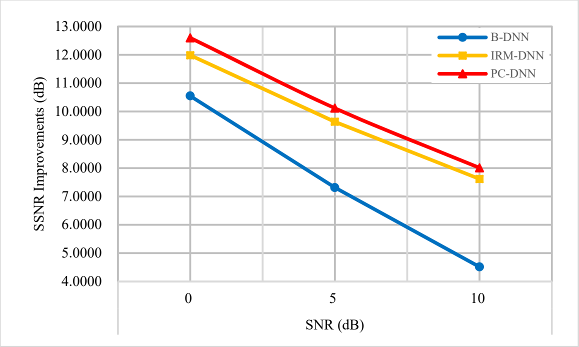

Fig. 4 The average SSNR improvements

The PESQ scores clearly reflect the test results of the three methods. We observe that although the B-DNN method achieves greatly improvement comparing with the noisy speech, the scores of the IRM-DNN method and the PC-DNN method are still superior to the directly mapped B-DNN method when the training set is the same. Furthermore, it can be easily seen that the PC-DNN method has a better improvement than the IRM-DNN method at all SNR levels, even *Factory* and *Street* noises are not contained in the training set. Comparing to B-DNN and IRM-DNN, which are already near perfect methods, the increase in PESQ scores of PC-DNN is due to accurate estimation of noise and the usage of phase correction. They help the enhanced speech to have more accurate phase information for the reconstruction of speech.

Fig. 3 shows ESTOI comparison of three methods for four noises under three input SNRs. From Fig. 3, an obvious improvement can be observed for the ESTOI results. We find that although the intelligibility of B-DNN and IRM-DNN is improved significantly for the noisy speech, the PC-DNN method still shows a better performance on the intelligibility tests, even for the two noises of non-training (*Factory* and *Street*). This indicates that the enhanced speech of the PC-DNN method has better speech quality than the B-DNN and IRM-DNN method under the three SNR levels (0dB, 5dB, 10dB).

Fig. 4 shows the SSNR comparison results corresponding to the improvements with the noisy speech. We can observe that for different SNR levels, the B-DNN and IRM-DNN methods have a very good improvement effect compared to noisy speech in the average SSNR improvements. The PC-DNN also has a certain amount of improvement in SSNR measurement through the joint estimation of clean speech, noise and the application of phase correction techniques, even if there are two types of test noise (*Factory* and *Street*) is not included in the training set.

As can be seen from the above experiments, we have obtained more accurate magnitude and phase information, which can recover a higher quality clean speech, by using DNN-based multi-task learning method and phase compensation technique. Both subjective and objective evaluations have good results.

IV. Conclusions

In this paper, we proposed a new method for speech enhancement, which is based on the DNN and phase correction strategy. The proposed method is IRM-DNN-based multi-task learning method of estimating magnitude spectrum of the clean speech and noise. We used the phase compensation function to correct the phase of the noisy speech to obtain more accurate phase spectral features. In comparison with the conventional DNN-based speech enhancement methods, the proposed method can significantly improve the enhanced speech quality and intelligibility. In the future work, we will further improve the accuracy of noise estimation and phase correction method, and focus on achieving the speech enhancement at much lower SNR levels.

Acknowledgment

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

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, no. 2, pp. 113-120, Apr 1979.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," in Proceedings of the IEEE, vol. 67, no. 12, pp. 1586-1604, Dec. 1979.J. Clerk Maxwell, *A Treatise on Electricity and Magnetism, 3*rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] F. Asano, S. Hayamizu, T. Yamada and S. Nakamura, "Speech enhancement based on the subspace method," in IEEE Transactions on Speech and Audio Processing, vol. 8, no. 5, pp. 497-507, Sep 2000.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral magnitude estimator," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 6, pp. 1109-1121, Dec 1984.

[5] P. Mowlaee, R. Saeidi, and Y. Stylianou, "INTERSPEECH 2014 Special Session: Phase Importance in Speech Processing Applications",Proc. Interspeech, pp. 1623-1627, 2014.

[6] T. Gerkmann, M. Krawczyk-Becker and J. Le Roux, "Phase Processing for Single-Channel Speech Enhancement: History and recent advances," in IEEE Signal Processing Magazine, vol. 32, no. 2, pp. 55-66, March 2015.

[7] Y. Wang, A. Narayanan and D. Wang, "On Training Targets for Supervised Speech Separation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 1849-1858, Dec. 2014.

[8] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, 2015, pp. 4390-4394.

[9] Wang, De Liang, and J. Chen. "Supervised Speech Separation Based on Deep Learning: An Overview." , 2017.

[10] Y. Xu, J. Du, L. R. Dai and C. H. Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7-19, Jan. 2015.

[11] Kingma, Diederik P, and J. Ba. "Adam: A Method for Stochastic Optimization." , Computer Science, 2014.

[12] P. S. Huang, M. Kim, M. Hasegawa-Johnson and P. Smaragdis, "Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 12, pp. 2136-2147, Dec. 2015.

[13] PALIWAL K, WÓJCICKI K, SHANNON B. "The Importance of Phase in Speech Enhancement". Speech Communication, vol.53(4), pp.465-494, 2011.

[14] Z. Wenlu and P. Hua, "Modified Wiener filtering speech enhancement algorithm with phase spectrum compensation," 2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN), Guangzhou, 2017, pp. 1075-1079.

[15] Wei Han, Xiongwei Zhang, Gang Min and Xingyu Zhou, "A novel single channel speech enhancement based on joint Deep Neural Network and Wiener Filter," 2015 IEEE International Conference on Progress in Informatics and Computing (PIC), Nanjing, 2015, pp. 163-167.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n,* vol. 93, 1993.

[17] G.H, "100 nonspeech environmental sounds," 2014.

[18] 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, no. 3, pp. 247-251, 1993

[19] 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 Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on, 2001, pp. 749-752.

[20] E J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibilityof speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*,vol. 24, no. 11, pp. 2009–2022, Nov 2016.

[21] John H L, Bryan H, Pellom L," An Effective Quality Evaluation Protocol For Speech Enhancement Algorithms," International Conference on Speech & Language Processing, 1988, pp. 2819-2822.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, no. 7, pp. 2125-2136, 2011.

[23] D. S. Williamson, Y. Wang and D. Wang, "Complex Ratio Masking for Monaural Speech Separation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 3, pp. 483-492, March 2016.

[24] M. Krawczyk and T. Gerkmann, "STFT Phase Reconstruction in Voiced Speech for an Improved Single-Channel Speech Enhancement," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 1931-1940, Dec. 2014.

[25] Yi Hu and P. C. Loizou, "A perceptually motivated approach for speech enhancement," in IEEE Transactions on Speech and Audio Processing, vol. 11, no. 5, pp. 457-465, Sept. 2003.