Deep Denoising Autoencoder Based Post Filtering for Speech Enhancement

Ryandhimas E. Zezario*, Jen-Wei Huang†, Xugang Lu§, Yu Tsao*, Hsin-Te Hwang‡, Hsin-Min Wang‡
*Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
Email: {ryandhimas, yu.tsao}@citi.sinica.edu.tw
†Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C.
Email: jwhuang@mail.ncku.edu.tw
§National Institute of Information and Communications Technology, Japan
Email: xugang.lu@nict.go.jp
‡Institute of Information Science, Academia Sinica, Taipei, Taiwan
Email: {hwanght, whm}@iis.sinica.edu.tw

Abstract— **In this paper, we present a simple yet effective deep denoising autoencoder (DDAE) based post-filter (DPF) approach for speech enhancement (SE). The DPF is designed to estimate the spectral difference of clean-noisy speech pair based on the enhanced-noisy speech pair. The difference estimated by the DPF approach is then used to compensate the noisy speech to obtain the final enhanced speech. We integrate the proposed DPF approach with one traditional SE method (minimum mean square error) and one deep-learning-based SE method (DDAE). Experiments on various noise types and signal-to-noise-ratio conditions were carried out to test the integrated systems. Results of three standardized objective evaluation metrics and automatic speech recognition (ASR) tests confirm that integrating the proposed DPF can improve the performance in further reducing spectral distortions and enhancing the speech quality and intelligibility.**

I. INTRODUCTION

Speech enhancement (SE) aims to retrieve clean speech signals from noise-corrupted ones and has served as a key unit in various speech-related applications, such as cochlear implants [1], hearing aids [2], automatic speech recognition (ASR) systems [3], and voice over internet protocol [4]. Traditionally, SE methods were derived based on some characteristics and statistical assumptions of clean speech and noise signals. Notable approaches include spectral subtraction [5], Karhunen–Loeve transform [6], Wiener filter [7], and minimum mean square error (MMSE) [8]. In recent years, machine-learning-based algorithms have been introduced to the SE field. Different from the traditional methods, a machine-learning-based SE approach generally prepares a denoising model in a data-driven manner without imposing strong statistical constraints. Well-known machine-learning-based models include nonnegative matrix factorization [9], compressive sensing [10], extreme learning machine [11], and deep learning models [12-21].

More recently, the locally linear embedding (LLE) algorithm, a well-known manifold learning algorithm, has been used to design a post-filter to further improve the enhanced speech processed by both traditional and machine-learning-based SE methods [22, 23]. Two LLE-based post-filter approaches, namely directly mapping and difference compensation, have been derived. Experimental results in [22, 23] first confirmed that both post-filter approaches could further improve the speech quality and intelligibility of the enhanced speech processed by MMSE [8] and deep denoising autoencoder (DDAE) [15] SE methods. Moreover, the difference compensation approach yielded better performance than the directly mapping counterpart. On the basis of the success of the LLE-based post-filter strategy, the present study proposes to adopting a deep learning based model, i.e., the DDAE model, as the difference compensation post-filter, termed DPF, to perform SE.

The DPF approach consists of offline and online stages. In the offline stage, the spectral difference of {enhanced speech; noisy speech} (termed DEN) is used as the input feature while the spectral difference of {clean speech, noisy speech} (termed DCN) is used as the output feature to train the DPF model. More specifically, the DPF model aims to characterize the mapping function of transforming DEN features to DCN features. In the online stage, we first compute the DEN features based on the spectral differences of {enhanced speech; noisy speech} of the testing speech. Next, by inputting the DEN features into the DPF model, we can obtain the predicted DCN features. Finally, the enhanced speech is obtained by compensating the spectral features of the noisy speech with the predicted DCN features. In this study, we investigate the compatibility of the DPF approach with one traditional SE method (MMSE) and one deep-learning-based SE method (DDAE).

To evaluate the proposed DPF approach, we adopted the Mandarin hearing in noise test (MHINT) sentences [24], which consisted of 300 utterances pronounced by a male native Mandarin speaker recorded in a clean condition room. Three metrics for objective evaluations were adopted: perceptual evaluation of speech quality (PESQ) [25], short-time objective intelligibility measure (STOI) [26], and log-spectral distortion (LSD) [27]. Besides, we tested ASR performance using the DPF enhanced speech. Experimental results confirm the effectiveness of the DPF approach in obtaining notable lower LSD scores, higher PESQ and STOI scores, and better ASR results, as compared to the corresponding single-stage SE systems.

The remainder of this paper is organized as follows. Section 2 presents the DPF approach. Section 3 presents the experimental setup and results. Section 4 concludes this work.

II. THE PROPOSED OF DPF APPROACH

Figures 1 and 2, respectively, show the offline and online stages of the DPF, which will be detailed in this section.

A. The Offline Stage

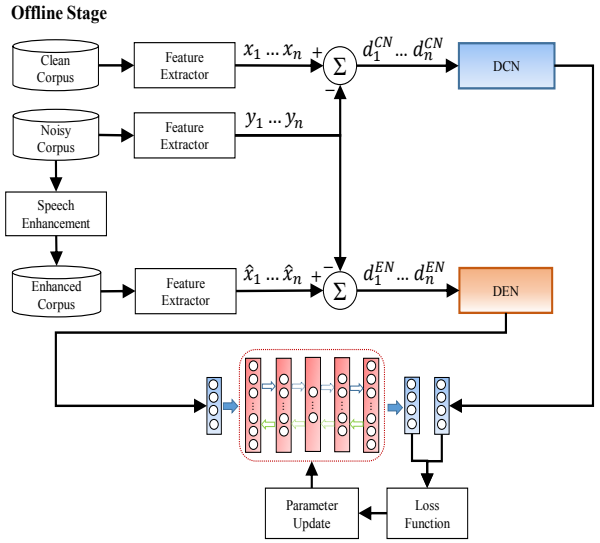

Fig. 1 *The offline stage of the DPF approach.*

In the offline stage, as shown in Fig. 1, the spectral difference of {enhanced speech, noisy speech} (DEN) and the spectral difference of {clean speech, noisy speech} (DCN) are first calculated. Suppose that we have clean speech feature vectors ($\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \dots, \boldsymbol{x}_N]$), noisy speech feature vectors ($\boldsymbol{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \dots, \boldsymbol{y}_N]$) and enhanced speech feature vectors ($\widehat{\boldsymbol{X}} = [\hat{\boldsymbol{x}}_1, \dots, \hat{\boldsymbol{x}}_n \dots, \hat{\boldsymbol{x}}_N]$), where $\boldsymbol{x}_n$, $\boldsymbol{y}_n$, and $\hat{\boldsymbol{x}}_n$ are the n-th clean, noisy, and enhanced vectors, respectively; N is the total number of training samples. The DEN features are computed by $\boldsymbol{D}^{EN} = \widehat{\boldsymbol{X}} - \boldsymbol{Y} = [\boldsymbol{d}_1^{EN}, \dots, \boldsymbol{d}_n^{EN}, \dots, \boldsymbol{d}_N^{EN}]$, where $\boldsymbol{d}_n^{EN} = \hat{\boldsymbol{x}}_n - \boldsymbol{y}_n$, and the DCN features are computed by $\boldsymbol{D}^{CN} = \boldsymbol{X} - \boldsymbol{Y} = [\boldsymbol{d}_1^{CN}, \dots, \boldsymbol{d}_n^{CN}, \dots, \boldsymbol{d}_N^{CN}]$, where $\boldsymbol{d}_n^{CN} = \boldsymbol{x}_n - \boldsymbol{y}_n$.

With the DEN and DCN feature sets, the DPF approach estimates the mapping function $F(\cdot)$:

$$\boldsymbol{d}_n^{CN} = F(\boldsymbol{d}_n^{EN}). \tag{1}$$

In this study, the DDAE model is used to model the mapping function, $F(\cdot)$, and thus, Eq. (1) can be re-written as:

$$\begin{aligned} h^1(\boldsymbol{d}_n^{EN}) &= \sigma(\boldsymbol{W}^0 \boldsymbol{d}_n^{EN} + \boldsymbol{b}^0), \\ &\vdots \\ h^J(\boldsymbol{d}_n^{EN}) &= \sigma(\boldsymbol{W}^{J-1} h^{J-1}(\boldsymbol{d}_n^{EN}) + \boldsymbol{b}^{J-1}), \\ \widehat{\boldsymbol{d}}_n^{CN} &= \boldsymbol{W}^J h^J(\boldsymbol{d}_n^{EN}) + \boldsymbol{b}^J, \end{aligned} \tag{2}$$

where $\{\boldsymbol{W}^0 \dots \boldsymbol{W}^J\}$ are the weight matrices, $\{\boldsymbol{b}^0 \dots \boldsymbol{b}^J\}$ are the bias vectors, and $\widehat{\boldsymbol{d}}_n^{CN}$ is the computed output given the input $\boldsymbol{d}_n^{EN}$; $\sigma(.)$ is an activation function, and the sigmoid function is used in this study. The parameter set of the DDAE model are estimated by:

$$\begin{aligned} \theta^* = \arg\min_{\theta} (L(\theta) + \eta^0 \|\boldsymbol{W}^0\|_{\mathrm{F}}^2 + \cdots + \\ \eta^J \|\boldsymbol{W}^J\|_{\mathrm{F}}^2), \end{aligned} \tag{3}$$

where

$$L(\theta) = \frac{1}{N} \sum_{n=1}^{N} \left\| \boldsymbol{d}_n^{CN} - \widehat{\boldsymbol{d}}_n^{CN} \right\|_2^2, \tag{4}$$

where $\theta = \{\boldsymbol{W}^0 \dots \boldsymbol{W}^J; \boldsymbol{b}^0 \dots \boldsymbol{b}^J\}$ is the parameter set of DDAE. In Eq. (3), $\{\eta^0 \dots \eta^J\}$ controls the tradeoff between the reconstruction accuracy and regularization of the weighting coefficients, and $\|.\|_{\mathrm{F}}^2$ denotes the Frobenius norm. We set $\eta^0 = \cdots = \eta^J$=0.0002 and use a Hessian-free algorithm to compute the parameters of the DDAE model, θ.

B. The Online Stage

In the online stage, given the noisy speech features $\overline{\boldsymbol{Y}} = [\overline{\boldsymbol{y}}_1, \dots, \overline{\boldsymbol{y}}_m \dots, \overline{\boldsymbol{y}}_M]$, where M denotes the total number of frames in the testing utterance, we first obtain the enhanced speech features, $\overline{\overline{\boldsymbol{X}}} = [\overline{\overline{\boldsymbol{x}}}_1, \dots, \overline{\overline{\boldsymbol{x}}}_m \dots, \overline{\overline{\boldsymbol{x}}}_M]$. Then, we compute the DEN features $\overline{\boldsymbol{D}}^{EN} = [\overline{\boldsymbol{d}}_1^{EN}, \dots, \overline{\boldsymbol{d}}_m^{EN}, \dots, \overline{\boldsymbol{d}}_M^{EN}]$, where $\overline{\boldsymbol{d}}_m^{EN} = [\overline{\overline{\boldsymbol{x}}}_m - \overline{\boldsymbol{y}}_m]$. Based on the computed DEN features, we then estimate the predicted DCN features using the DDAE model that is trained in the offline stage:

$$\overline{\boldsymbol{d}}_m^{CN} = F(\overline{\boldsymbol{d}}_m^{EN}). \tag{5}$$

Then $\overline{\boldsymbol{d}}_m^{CN}$ is used to perform feature compensation:

$$\overline{\boldsymbol{x}}_m = \overline{\boldsymbol{y}}_m + \overline{\boldsymbol{d}}_m^{CN}, \tag{6}$$

where $\overline{\boldsymbol{x}}_m$ is the final enhanced speech. The phase of the noisy speech is used as the phase of the enhanced speech. Finally, an inverse FFT (IFFT) is applied to convert the enhanced spectral and phase features to obtain the enhanced speech.

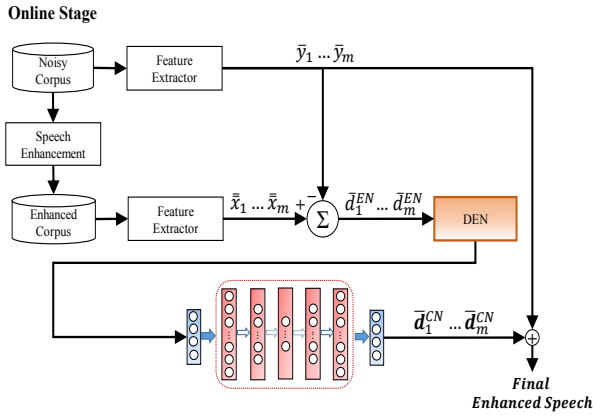

Fig. 2 *The online stage of the DPF approach.*

III. EXPERIMENTS

This section first presents the experimental setup and then demonstrates the experimental results. We intend to investigate the compatibility of DPF with different types of SE methods, and thus MMSE (a traditional statistical-model-based SE method) and DDAE (a deep-learning-based SE method) have been adopted as the first-stage SE methods to generate the initially enhanced speech, which is then processed by the DPF approach, as introduced in Section 2.

A. Experimental Setup

As mentioned in Section 1, the MHINT sentences [24] were used to test the proposed DPF approach. Among the 300 clean speech utterances in the MHINT sentences, 250 utterances were used as the training data, and the remaining 50 utterances were used as the testing data. Two types of noises, car and two-talker recorded in real environments, were used to generate noisy speech. All of the speech and noise signals were recorded in a 16 kHz/16 bit format. We prepared noisy speech by artificially adding noises to the clean speech. For the training set, the signal-to-noise ratio (SNR) ranged from -10 to 20 dB (with a 5 dB interval). As a result, for each noise type, 1750 noisy speech utterances (5×250) along with the corresponding clean speech utterances were prepared as the training set. For the testing set, we prepared noisy speech with the SNR ranged from -10 to 10 dB (with a 4 dB interval) for both noise types.

To train the DPF model, we first obtained 1750 enhanced speech utterances by applying the first-stage SE methods on the 1750 noisy speech in the training set. With the enhanced and noisy speech, we can prepare the DEN features (the input training samples of the DPF model). Meanwhile, based on the 1750 noisy speech utterances and their corresponding clean versions, we can prepare the DCN features (the output training samples of the DPF model). The DPF model was realized by a DDAE model consisting of three hidden layers, with 2500 hidden nodes in each layer. For signal analysis, the frame length and the frame shift were 32 and 16 milliseconds, respectively. The Hamming window was used in the framing process. Each frame of speech signals was further converted to a 257-dimensional log power spectral feature vector.

We evaluated the performance of integrating the DPF approach with the MMSE and DDAE SE methods. In the following discussions, the single-stage MMSE and DDAE SE methods are denoted as MMSE and DDAE, respectively, and the integrated DPF with MMSE and DDAE are denoted as DDAE-DPF and MMSE-DPF, respectively.

B. Objective Evaluation

The PESQ [25], STOI [26], and LSD [27] metrics were used as the signal-level evaluation measures. The PESQ metric indicates the speech quality with a score ranging from -0.5 to 4.5. The STOI metric indicates the speech intelligibility, with a score ranging from 0 to 1. The LSD metric indicates the difference between the enhanced and clean speech. For PESQ and STOI, a higher score stands for better speech quality and intelligibility, respectively. For LSD, a lower score denotes that the enhanced speech is closer to the clean speech.

We first investigated the performance of integrating DPF with MMSE. The objective scores of MMSE and MMSE -DPF are listed in Tables 1 and 2, respectively, for the car and two-talker noisy conditions. The minima controlled recursive averaging noise tracking algorithm [28] was used to compute the required statistics for the MMSE SE method. From the two tables, consistent improvements of PESQ, STOI, and LSD were noted in both car and two-talker noises for all SNR levels when the DPF was applied, confirming the effectiveness of the DPF approach to further enhance speech quality and intelligibility over the single-stage MMSE SE method.

Next, we investigated the performance of integrating DPF with DDAE. Tables 3 and 4 demonstrate the results of the DDAE and DDAE-DPF in the car and two-talker noisy conditions, respectively. We observed the same trends as those from Tables 1 and 2: DDAE-DPF outperforms DDAE in terms of PESQ, STOI, and LSD metrics consistently over all SNR levels for both noisy conditions. Please note that for DDAE-DPF, both the first-stage enhancement and the post-filter processing were conducted by the DDAE model with the same training criteria. The only difference of the first-stage enhancement and the post-filter processing was the input-output of the DDAE model. For the first-stage enhancement, the input-output were noisy-clean spectral features, while for the post-filter, the input-output were DEN-DCN features (spectral differences). Note that we adopted the log power spectral features as the spectral features in this study. Therefore, the spectral difference is actually the spectral ratio in the linear domain. In other words, the first-stage DDAE model performed direct spectral mapping, and the second DDAE model (serves as a DPF) predicted the ratio of clean to noisy speech in order to compute the enhanced speech. It has been reported in [29] that direct spectral mapping and ratio-masking mechanisms can be used together to leverage the complementary information to achieve better SE performance. The integrated DDAE SE model and DPF (DDAE-based post-filter) presented in this study can be considered as a cascading approach to combine the knowledge of spectral mapping and the ratio-masking mechanisms when performing SE.

TABLE 1.
PESQ, STOI, AND LSD OF MMSE AND MMSE-DPF IN THE CAR NOISY CONDITION.

	PESQ		STOI		LSD	
Method	MMSE	MMSE-DPF	MMSE	MMSE-DPF	MMSE	MMSE-DPF
SNR 10	3.13	**3.17**	0.92	**0.95**	1.10	**0.79**
SNR 6	2.79	**2.86**	0.89	**0.92**	1.12	**0.93**
SNR 2	2.43	**2.49**	0.85	**0.88**	1.18	**1.09**
SNR 0	2.27	**2.33**	0.82	**0.86**	1.21	**1.18**
SNR -2	2.11	**2.14**	0.79	**0.84**	**1.26**	**1.26**
SNR -6	1.83	**1.84**	0.74	**0.79**	**1.40**	1.44
SNR -10	1.58	**1.61**	0.67	**0.74**	**1.55**	1.58
Ave	2.31	**2.35**	0.81	**0.86**	1.26	**1.18**

TABLE 2.
PESQ, STOI, AND LSD OF MMSE AND MMSE-DPF IN THE TWO- TALKER NOISY CONDITION.

	PESQ		STOI		LSD	
Method	MMSE	MMSE-DPF	MMSE	MMSE-DPF	MMSE	MMSE-DPF
SNR 10	2.16	**2.33**	**0.89**	**0.89**	1.42	**1.10**
SNR 6	1.84	**1.97**	0.83	**0.85**	1.57	**1.27**
SNR 2	1.56	**1.73**	0.74	**0.79**	1.75	**1.46**
SNR 0	1.50	**1.63**	0.70	**0.75**	1.85	**1.53**
SNR -2	1.42	**1.52**	0.64	**0.71**	1.92	**1.63**
SNR -6	1.30	**1.35**	0.52	**0.64**	2.10	**1.76**
SNR -10	1.22	**1.26**	0.42	**0.54**	2.24	**1.95**
Ave	1.57	**1.68**	0.68	**0.74**	1.83	**1.53**

TABLE 3.
PESQ, STOI, AND LSD OF DDAE AND DDAE-DPF IN THE CAR NOISY CONDITION.

	PESQ		STOI		LSD	
Method	DDAE	DDAE-DPF	DDAE	DDAE-DPF	DDAE	DDAE-DPF
SNR 10	2.72	**3.32**	0.89	**0.95**	0.77	**0.61**
SNR 6	2.59	**3.04**	0.88	**0.93**	0.81	**0.69**
SNR 2	2.37	**2.71**	0.86	**0.90**	0.89	**0.78**
SNR 0	2.24	**2.53**	0.85	**0.88**	0.93	**0.83**
SNR -2	2.16	**2.39**	0.84	**0.86**	0.96	**0.88**
SNR -6	1.95	**2.13**	0.80	**0.82**	1.09	**1.02**
SNR -10	1.74	**1.88**	0.77	**0.78**	1.22	**1.16**
Ave	2.25	**2.57**	0.84	**0.87**	0.95	**0.85**

TABLE 4.
PESQ, STOI, AND LSD OF DDAE AND DDAE-DPF IN THE TWO- TALKER NOISY CONDITION.

	PESQ		STOI		LSD	
Method	DDAE	DDAE-DPF	DDAE	DDAE-DPF	DDAE	DDAE-DPF
SNR 10	2.41	**3.06**	0.88	**0.93**	0.87	**0.76**
SNR 6	2.21	**2.73**	0.86	**0.91**	0.92	**0.83**
SNR 2	1.99	**2.47**	0.84	**0.88**	0.96	**0.89**
SNR 0	1.91	**2.33**	0.82	**0.86**	**1.00**	1.16
SNR -2	1.81	**2.18**	0.81	**0.84**	1.04	**0.93**
SNR -6	1.60	**1.92**	0.76	**0.80**	1.14	**1.03**
SNR -10	1.44	**1.69**	0.71	**0.74**	1.31	**0.96**
Ave	1.91	**2.34**	0.81	**0.85**	1.04	**0.94**

C. *ASR Performance*

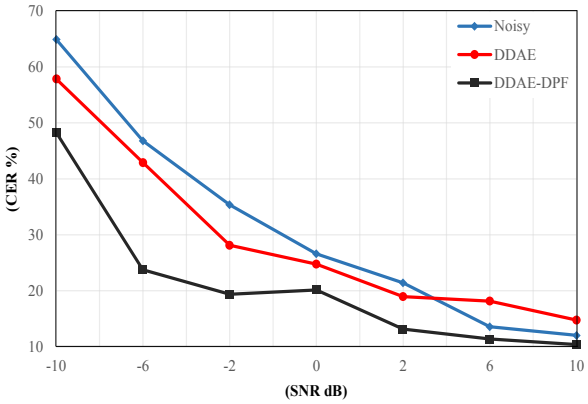

Fig. 3: *CER of Google ASR for noisy speech, DDAE, and DDAE-DPF.*

The results from the previous experiments have confirmed that the proposed DPF model can yield higher scores in three signal-level objective measures, namely PESQ, STOI, and LSD. Here we test the ASR performance using the enhanced speech to further confirm the effectiveness of the DPF approach. We used the Google ASR [30] as the speech recognizer, considering that ASR systems are often developed by a third-party in most real world senarios. When testing recognition, the DPF enhanced speech was used as the input to the Google ASR,

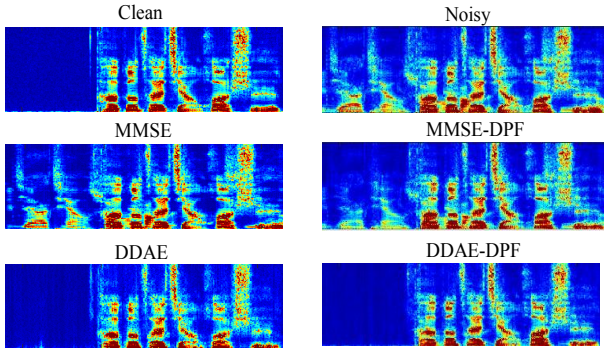

Fig. 4: *Spectrograms of a clean utterance, with its noisy, MMSE, DDAE, MMSE-DPF, and DDAE-DPF versions in the two-talker noise at 6dB SNR condition.*

and the character error rate (CER) in Mandarin was commuted based on the correct transcription reference. The CER results of the car noise were reported in Fig. 3. In addition to DDAE-DPF, the results of the unprocessed noisy speech (denoted as Noisy) and the single-stage DDAE enhanced speech (denoted as DDAE) are also presented for comparison.

From the results in Fig. 3, we can note that when comparing to the unprocessed noisy speech, the speech processed by the single-stage DDAE achieved lower CERs in relatively noisier conditions (-10 to 2 dB SNR levels) while higher CERs in relatively cleaner conditions (SNR higher than 6 dB). On the other hand, DDAE-DPF achieves further improvements over the single-stage DDAE and outperforms unprocessed noisy speech consistently over low to high SNR levels.

D. *Spectrogram Analysis*

In addition to the objective measures and ASR tests, we also present the spectrogram plots in order to visually investigate the charactersitcs of the DPF enhanced speech. Figure 4 shows the spectrograms of the clean and noisy speech utterances at 6 dB SNR level under two-talker noise. The spectrograms of the MMSE and DDAE enhanced speech utterances along with those further improved by the DPF are presented below the clean and noisy speech spectrgrams. From the figure, we can note that almost all of the SE methods can effectively reduce noise components from the noisy speech utterance. We also observe that the DPF approach can further improve the MMSE and DDAE enhanced speech by eliminating distortions and restoring the detailed information of the speech signals. From the objective evaluation results reported in Tables 1 to 4 and the spectrograms shown in Fig. 4, we can note that DDAE-DPF yields better enhancement results than MMSE-DPF. Since the single-stage DDAE can achieve better performance than the single-stage MMSE, the results suggest that the overall performance of the DPF approach depends on the capability of the preceding SE method, which has also been noted in [22, 23].

IV. CONCLUSION

This paper has presented a novel post-filter approach (DPF) based on the DDAE model. Experimental results show that the DPF approach can further improve both traditional and deep-

learning-based SE methods in terms of signal-level objective evaluations (PESQ, STOI, and LSD) and ASR tests. The major contribution of this paper is that the DDAE model, which has been confirmed to provide satisfactory performance in spectral mapping, can also be used as a post-filter to further improve the enhanced speech. Moreover, the improvements provided by DDAE-DPF over the single-stage DDAE confirmed the advantage of combining the knowledge of spectral mapping and the ratio-masking mechanisms to achieve better SE performance. In the future, we will explore the applicability of DPF in different noise types, the compatibility with other SE methods, and the unseen condition of test data.

V. ACKNOWLEDGEMENT

The authors would like to thank the Ministry of Science and Technology for providing financial supports (105-2221-E-006 -212 -MY2, 106-2221-E-001-017-MY2, and 107-2221-E-001-012-MY2).

REFERENCES

[1] Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1568-1578, 2017.

[2] D. Wang, "Deep learning reinvents the hearing aid," *IEEE Spectrum*, March Issue, pp. 32-37 (Cover Story), 2017.

[3] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, "Robust automatic speech recognition: a bridge to practical applications," *1st ed. Academic Press*, 2015.

[4] S. Han, S. Jeong, H. Yang, J. Kim, "Noise reduction for VoIP speech codecs using modified Wiener filter," *Innovations in Systems, Computing Sciences and Software Engineering*, p.393–397, 2007.

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

[6] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Trans. Speech Audio Process*., vol. 9, no. 2, pp. 87–95, Feb. 2001.

[7] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," *in Proc. ICASSP*, pp. 629–632, 1996.

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

[9] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," *in Proc. ICASSP*, pp. 4029–4032, 2008.

[10] J.-C. Wang, Y.-S. Lee, C.-H. Lin, S.-F. Wang, C.-H. Shih, and C.-H. Wu, "Compressive sensing-based speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 24, no. 11, pp. 2122 – 2131, 2016.

[11] T. Hussain, S. M. Siniscalchi, C.-C. Lee, S.-S. Wang, Y. Tsao, W.-H. Liao, "Experimental study on extreme learning machine applications for speech enhancement," *IEEE Access,* vol. 5, pp. 25542 – 25554, 2017.

[12] D. Wang and J. Chen, "Supervised speech separation based on deep learning: an overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702-1726, 2018.

[13] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[14] 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, no. 1, pp. 65–68, 2014

[15] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," *in Proc. INTERSPEECH*, pp. 436–440, 2013

[16] B. Xia, C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Commun*, 60, 13-29, 2014.

[17] M. Kolbak, Z.-H. Tan, J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE Trans Audio Speech Language Process*, 25, 153-167, 2017.

[18] F. Weninger, H. Erdogan, S.Watanabe, et al, " Speech enhancement with LSTM recurrent neural networks and its application to noise-rebust ASR," *International conference on latent variable anaysis and signal separation*, pp. 91-99, 2015.

[19] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-Aware convolutional neural network modeling for speech enhancement," *in Proc. INTERSPEECH*, pp. 8–12, 2016.

[20] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *APSIPA*, 2017.

[21] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol 26, no. 9, pp.1570-1584, 2018.

[22] Y.-C. Wu, H.-T. Hwang, S.-S. Wang, C.-C. Hsu, Y.-H. Lai, Y. Tsao, and H.-M. Wang, "A locally linear embedding based postfiltering approach for speech enhancement," *in Proc. ICASSP*, 2017.

[23] H.-T. Hwang, Y.-C. Wu, S.-S. Wang, C.-C. Hsu, Y. Tsao, H.-M. Wang, Y.-R. Wang, and S.-H. Chen, "Locally linear embedding based post-filtering for speech enhancement," *to appear in Journal of Information Science and Engineering*.

[24] L. L. Wong, S. D. Soli, S. Liu, N. Han, and M.-W. Huang, "Development of the mandarin hearing in noise test (MHINT)," *Ear and Hearing*, vol. 28, no. 2, pp. 70S–74S, 2007

[25] ITU-T, Rec. P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *International Telecommunication Union-Telecommunication Standardisation Sector*, 2001.

[26] 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.

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

[28] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, 2002.

[29] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple target deep learning for LSTM-RNN based speech enhancement," *in Proc. HSCMA*, pp. 136–140, 2017.

[30] A. Zhang. (2017). *Speech Recognition (Version 3.6) [Software]*. Available: https://github.com/Uberi/speech_recognition#readme