Speech dereverberation using weighted prediction error with prior learnt from data

Ziye Yang, Wenxing Yang, Kai Xie, Jie Chen

Research and Development Institute of Northwestern Polytechnical University in Shenzhen, China
CIAIC, School of Marine Science and Technology, Northwestern Polytechnical University, China
{yzy97,yangwenxing521,xiekai}@mail.nwpu.edu.cn, dr.jie.chen@ieee.org

***Abstract*—Speech dereverberation aims to mitigate the impact of late-reverberant components. As a typical approach to dereverberation, the weighted prediction error (WPE) method has shown its superior performance, however it is still possible to further improve its performance and robustness by incorporating sophisticated speech priors. Recent research demonstrates that the integration of physics-based and data-driven methods can improve the performance of various signal processing tasks while maintaining the interpretability of the problem solving process. Motivated by the relevant progress, this paper presents a novel dereverberation framework that incorporates the data-driven method for speech prior capturing for WPE. The plug-and-play strategy (PnP), specifically the regularization by denoising (RED) strategy, is used to incorporate speech prior information during the alternating direction method of multipliers (ADMM) solving iterations by plugging in a pre-trained speech denoiser. Experimental results demonstrate the effectiveness of the proposed method[1].**

***Index Terms*—Speech dereverberation, the weighted prediction error method, data-driven method, learnt speech priors**

I. Introduction

The speech signals captured by microphones in an enclosed room unavoidably contain reverberant components, resulting in a degradation of the quality of interested speech and further impairing the automatic speech recognition system. Therefore, speech dereverberation techniques have been widely investigated [1]–[3]. These techniques aim at eliminating the late-reverberant components and preserving the direct-path and early-reverberant components, since the former are detrimental to both speech intelligibility and quality [4], [5].

Extensive works have been devoted to devising speech dereverberation methods, which can be primarily divided into conventional physics-based and data-driven algorithms. The former usually solves the dereverberation problem based on the speech convolution model, possessing a clear physical interpretation. Among the numerous physics-based dereverberation techniques [6]–[8], the class of multichannel linear prediction methods [9] is, to our knowledge, the most promising. Built on this, the weighted prediction error (WPE) method [10] in particular shows its effectiveness in dereverberation. In order to further enhance the performance, the work in [11] and [12] propose to exploit speech sparsity in the time-frequency domain to incorporate an additional prior on the unknown variance.

Recent data-driven methods [13], [14] heavily rely on deep learning and have become a hotspot due to their superior capability to excavate high-level features. This kind of method aims to learn a mapping function from the input signals to the output clear speech with the speech priors embedding in the network parameters. However, it is a black-box and may lack physical interpretability and generalizability. Recently, the integration of physics-based methods and data-driven methods has received considerable attention in the signal processing community [15]–[17]. For instance, in [18] a deep neural network-based spectrum estimator is incorporated into the vanilla WPE to boost dereverberation performance. Although useful, this work does not extract structural information of the speech spectrum from data. Among several possible strategies, the plug-and-play technique (PnP), which plugs deep denoising algorithms as a module into the optimization iterations to capture data priors, has been successfully investigated for various tasks [19]–[21].

Inspired by this advance, we intend to establish a framework for speech dereverberation that benefits from both the physics-based model and data priors. Specifically, we formulate the prediction error minimization problem of WPE with an additional regularizer that is not explicitly handcrafted. In contrast to the vanilla WPE method and extensions [11], [12], [18] that do not take into account sophisticated speech priors, integrating speech prior information learnt from data is expected to be advantageous. To this end, the PnP strategy, specifically the regularization by denoising (RED) strategy [22], is used to incorporate speech prior information during the alternating direction method of multipliers (ADMM) [23] solving iterations using a pre-trained speech denoiser. Experimental results validate the proposed method and show its improvement in performance and robustness over WPE.

Notation. Normal font letters x and X denote scalars, and boldface small letters $\mathbf{x}$ denote column vectors. Boldface capital letters $\mathbf{X}$ represent matrices, and the operator $(\cdot)^{\top}$ and $(\cdot)^{\mathrm{H}}$ denote matrix transpose and conjugate transpose respectively.

The work was supported in part by Department of Nature Resources of Guangdong (GDNRC[2023]47), Shenzhen Science and Technology Program JCYJ20220530161606014, Guangdong Intl. Coop. Project 2022A050505-0020, NSFC Grant 62171380, Shaanxi Key Industrial Innovation Chain Project 2022ZDLGY01-02, and Xi'an Technology Industrialization Plan XA2020-RGZNTJ-0076.

[1]Demo results are available at https://github.com/yzy-nwpu/PNP_WPE-for-speech_dereverb.

II. Problem Formulation

We consider the signal model under the scenario where a microphone array with Q channels captures the convolved speech with additive noise. In time domain, the observed signal of q-th channel can be represented by:

$$x_q(t) = h_q(t) * s(t) + y_q(t), \tag{1}$$

where $h_q(t)$ is the acoustic impulse response between the source and the microphone, $s(t)$ is the source speech, $*$ denotes the linear convolution, and $y_q(t)$ is the zero-mean additive noise that is independent of $s(t)$. Signal model (1) can be approximated in short-time Fourier transform (STFT) domain by [10]:

$$X_q(n,k) = \sum_{j=0}^{J-1} H_q(j,k) S(n-j,k) + Y_q(n,k), \tag{2}$$

where n and k are the time-frame and frequency bin indices respectively, J denotes the order of $H_q(n,k)$ which is the $h_q(t)$ in STFT domain, $X_q(n,k)$, $S(n,k)$ and $Y_q(n,k)$ represent the counterparts of $x_q(t)$, $s(t)$ and $y_q(t)$ in the STFT domain respectively.

Considering the multichannel linear prediction dereverberation process, the desired speech can be estimated by:

$$\hat{S}(n,k) = X_{\mathrm{ref}}(n,k) - \overline{\mathbf{w}}^{\mathrm{H}}(k)\overline{\mathbf{x}}(n-D,k), \tag{3}$$

where $\overline{\mathbf{x}}(n-D,k)$ is constructed by stacking $\{[X_q(n-D,k), X_q(n-D-1,k), \cdots, X_q(n-D-L+1,k)]^{\top}\}_{q=1}^{Q}$ to form a vector of length $L_Q = Q \times L$ with L being the filter order and D being a predefined delay, $\overline{\mathbf{w}}^{\mathrm{H}}(k)$ is the filter weight vector of length L_Q, and $X_{\mathrm{ref}}(n,k)$ denotes the reference signal which can be randomly chosen at any microphone. The WPE method seeks the filter weight vector by optimizing the following cost function:

$$\mathcal{J}_{\mathrm{WPE}}\Big(\{\overline{\mathbf{w}}(k)\}_{k=1}^{K}\Big) = \sum_{k=1}^{K}\sum_{n=1}^{N} \frac{|\hat{S}(n,k)|^2}{\sigma(n,k)} + \log \pi\sigma(n,k), \tag{4}$$

with $\hat{S}(n,k)$ defined in (3), and $\sigma(n,k)$ is the estimate of the speech variance at frame n and frequency bin k. Since the prediction error $\hat{S}(n,k)$ is considered the desired signal, it is beneficial to introduce a regularization term to incorporate speech priors for $\hat{S}(n,k)$:

$$\mathcal{J}_{\mathrm{WPE\_Reg}}\Big(\{\overline{\mathbf{w}}(k)\}_{k=1}^{K}\Big) = \mathcal{J}_{\mathrm{WPE}}\Big(\{\overline{\mathbf{w}}(k)\}_{k=1}^{K}\Big) + \beta\mathcal{J}_{\mathrm{Reg}}\Big(\hat{\mathbf{S}}\Big), \tag{5}$$

where β is a trade-off parameter, $\mathcal{J}_{\mathrm{Reg}}$ denotes a regularizer, and $\hat{\mathbf{S}}$ is the speech time-frequency matrix consisting of $\{\hat{S}(n,k)\}_{n,k=1}^{N,K}$ [2]. Designing a good regularizer J_{Reg} along with an efficient solving method is not a trivial task. Instead, we propose to learn priors from speech data and incorporate them into the mathematics-based optimization to address this problem based on the PnP strategy. Particularly, we consider $\mathcal{J}_{\mathrm{Reg}}$ in the form of:

$$\mathcal{J}_{\mathrm{Reg}}\Big(\hat{\mathbf{S}}\Big) = \frac{1}{2}\hat{\mathbf{S}}^{\mathrm{H}}\big[\hat{\mathbf{S}} - \Omega(\hat{\mathbf{S}})\big], \tag{6}$$

where $\Omega(\cdot)$ denotes an off-the-shelf denoiser. This form is called RED, which is an effective regularizer with favorable derivative properties under mild assumptions [22].

III. Solving Method and Network Design

In this section, we present the solving method for the problem defined by equations from (3) to (6), and the way to integrate data-driven speech priors with a denoising deep neural network.

A. Variable splitting based on ADMM

To solve the problem, we first introduce new variables $R(n,k)$ into the problem with additional equality constraints, leading to the following problem formulation:

$$\begin{aligned} \min_{\overline{\mathbf{w}}(k),\mathbf{R},\mathbf{V}} & \sum_{k=1}^{K}\sum_{n=1}^{N} \frac{|R(n,k)|^2}{\sigma(n,k)} + \log \pi\sigma(n,k) + \frac{\beta}{2}\mathbf{R}^{\mathrm{H}}\big[\mathbf{R} - \Omega(\mathbf{R})\big] \\ \text{s.t.} \quad & R(n,k) = X_{\mathrm{ref}}(n,k) - \overline{\mathbf{w}}^{\mathrm{H}}(k)\overline{\mathbf{x}}(n-D,k) - V(n,k) \\ & \mathcal{E}(\mathbf{R}) = \sigma_{\mathrm{norm}}^2, \end{aligned} \tag{7}$$

where $V(n,k)$ denotes the additive noise in the modeling and processing[3], $\mathcal{E}(\cdot)$ represents the energy of the argument signal, and σ_{norm}^2 denotes a predefined signal energy. Note that the energy constraint is introduced to ensure the uniqueness of the solution. Future extended manuscript will elaborate on the rationale behind the introduction of the noise term $V(n,k)$ and energy normalization in problem (7), which differs from the typical PnP procedure. The corresponding (scaled) augmented Lagrangian function is defined as:

$$\begin{aligned} & \mathcal{L}\Big(\overline{\mathbf{w}}(k)_{k=1}^{K}, \mathbf{R}, \mathbf{V}, \mathbf{P}\Big) \\ =& \mathcal{J}_{\mathrm{WPE}} + \frac{\beta}{2}\mathbf{R}^{\mathrm{H}}\big[\mathbf{R} - \Omega(\mathbf{R})\big] + \frac{\rho}{2}\sum_{k=1}^{K}\sum_{n=1}^{N}\Big(\big|[X_{\mathrm{ref}}(n,k) - \overline{\mathbf{w}}^{\mathrm{H}}(k) \\ & \times \overline{\mathbf{x}}(n-D,k)] - V(n,k) - R(n,k) + P(n,k)\big|^2 - |P(n,k)|^2\Big), \end{aligned} \tag{8}$$

where $P(n,k)$ is the scaled dual variable, and ρ is the penalty parameter. The ADMM decouples the optimization of (8) into solving subproblems over iteration index ℓ as follows.

1) Step 1 — Optimization with respect to $\overline{\mathbf{w}}(k)$: The optimization of (8) reduces to:

$$\begin{aligned} \overline{\mathbf{w}}^{(\ell+1)}(k) = \underset{\overline{\mathbf{w}}(k)}{\operatorname{argmin}} \, & \mathcal{J}_{\mathrm{WPE}} + \frac{\rho}{2}\sum_{k=1}^{K}\sum_{n=1}^{N}\big|X_{\mathrm{ref}}(n,k) \\ & - [\overline{\mathbf{w}}^{(\ell)}(k)]^{\mathrm{H}}\overline{\mathbf{x}}(n-D,k) - V^{(\ell)}(n,k) \\ & - R^{(\ell)}(n,k) + P^{(\ell)}(n,k)\big|^2. \end{aligned} \tag{9}$$

The optimization w.r.t. $\overline{\mathbf{w}}(k)$ is a separable least square problem and can be then solved by

$$\overline{\mathbf{w}}^{(\ell+1)}(k) = [R_{\overline{\mathbf{x}}}^{(\ell+1)}(k)]^{-1}\mathbf{p}_{\overline{\mathbf{x}}}^{(\ell+1)}(k), \tag{10}$$

[2]Note that the similar definition will be used for the other bold capital letters, such as $\mathbf{R}$, $\mathbf{V}$ and $\mathbf{P}$.

[3]Note that for the situation without noise, we can set $\mathbf{V}$ to 0 and simplify the optimization process, then $\mathbf{R}$ reduces to $\mathbf{S}$ in (5) and (6).

where

$$R_{\overline{\mathbf{x}}}^{(\ell+1)}(k) = \sum_{n=1}^{N} \frac{\overline{\mathbf{x}}(n-D,k)[\overline{\mathbf{x}}(n-D,k)]^{\mathrm{H}}}{\lambda^{(\ell+1)}(n,k)} \tag{11}$$

and

$$\mathbf{p}_{\overline{\mathbf{x}}}^{(\ell+1)}(k) = \sum_{n=1}^{N} \frac{\overline{\mathbf{x}}(n-D,k)\tilde{X}^{(\ell+1)}(n,k)}{\lambda^{(\ell+1)}(n,k)}. \tag{12}$$

In the above solution, $\lambda^{(\ell+1)}(n,k)$ is given by

$$\lambda^{(\ell+1)}(n,k) = \frac{2\sigma^{(\ell)}(n,k)}{2+\rho\sigma^{(\ell)}(n,k)}, \tag{13}$$

and $\tilde{X}^{(\ell+1)}(n,k)$ is given by

$$\begin{aligned}\tilde{X}^{(\ell+1)}(n,k) =& X_{\mathrm{ref}}(n,k) - \frac{\rho}{2}\lambda^{(\ell+1)}(n,k)\big[R^{(\ell)}(n,k) \\ &+ V^{(\ell)}(n,k) - P^{(\ell)}(n,k)\big].\end{aligned} \tag{14}$$

By substituting $\overline{\mathbf{w}}(k)$ of each band into (3), we can construct matrix $\hat{\mathbf{S}}$ which will be used in the following steps, and estimate $\sigma(n,k)$ based on $\hat{S}(n,k)$ as [10]:

$$\sigma^{(\ell+1)}(n,k) = |\hat{S}^{(\ell)}(n,k)|^2. \tag{15}$$

2) Step 2 — Optimization with respect to $\mathbf{R}$: The optimization problem (8) now reduces to

$$\begin{aligned}\mathbf{R}^{(\ell+1)} =& \underset{\mathbf{R}}{\operatorname{argmin}} \frac{\rho}{2}\left\|\hat{\mathbf{S}}^{(\ell+1)} - \mathbf{V}^{(\ell)} - \mathbf{R}^{(\ell)} + \mathbf{P}^{(\ell)}\right\|^2 \\ &+ \frac{\beta}{2}[\mathbf{R}^{(\ell)}]^{\mathrm{H}}\big[\mathbf{R}^{(\ell)} - \Omega(\mathbf{R}^{(\ell)})\big].\end{aligned} \tag{16}$$

From the perspective of RED [22], the prior properties of speech can be incorporated in (16) by applying a denoising processing to speech $\tilde{\mathbf{R}}^{(\ell+1)} = \hat{\mathbf{S}}^{(\ell+1)} - \mathbf{V}^{(\ell)} + \mathbf{P}^{(\ell)}$. The solution to this problem can be achieved via the fixed-point iteration:

$$\mathbf{R}^{(\ell+1,i)} = \mu\tilde{\mathbf{R}}^{(\ell+1,i)} + (1-\mu)\Omega(\tilde{\mathbf{R}}^{(\ell+1,i)};\Theta) \tag{17}$$

with $\mu = \frac{\rho}{\rho+\beta}$ and inner iteration $i = 1,\cdots,I$, where Θ denotes the parameters of the denoiser. Considering the energy normalization in (7), we also conduct the normalization

$$\mathbf{R}^{(\ell+1)} = \frac{\mathbf{R}^{(\ell+1)}}{\mathcal{E}(\mathbf{R}^{(\ell+1)})}\sigma_{\mathrm{norm}}^2, \tag{18}$$

where $\sigma_{\mathrm{norm}}^2 = \mathcal{E}(\hat{\mathbf{S}}^{(\ell+1)})$.

3) Step 3 — Optimization with respect to $\mathbf{V}$: Here, the solution of this optimization problem readily writes:

$$\mathbf{V}^{(\ell+1)} = \hat{\mathbf{S}}^{(\ell+1)} - \mathbf{R}^{(\ell+1)} + \mathbf{P}^{(\ell)}. \tag{19}$$

4) Step 4 — Update of $\mathbf{P}$: This dual variable is updated in the standard manner:

$$\mathbf{P}^{(\ell+1)} = \mathbf{P}^{(\ell)} + \hat{\mathbf{S}}^{(\ell+1)} - \mathbf{V}^{(\ell+1)} - \mathbf{R}^{(\ell+1)}. \tag{20}$$

Variables $\overline{\mathbf{w}}(k)$, $\mathbf{R}$, $\mathbf{V}$ and $\mathbf{P}$ are updated until convergence, and output $\mathbf{R}$ will be used as the estimated speech.

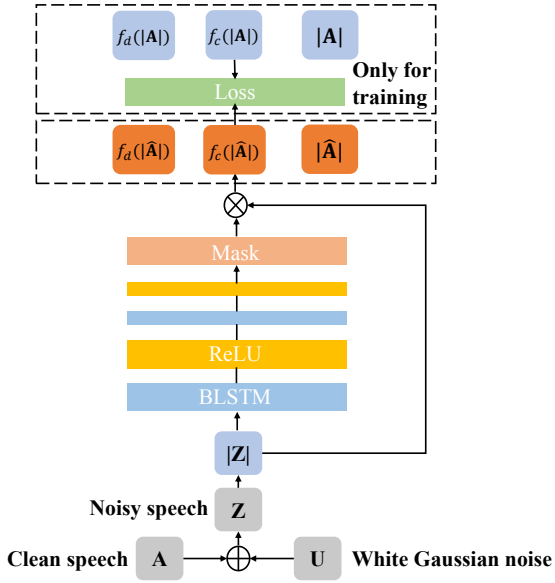


Fig. 1. Diagram of the BLSTM-based denoiser.

B. BLSTM-based denoiser

Any denoiser can be used for (17) to incorporate speech priors, making the proposed framework flexible. To focus on the main idea of this work, here we simply train a bidirectional long short-term memory (BLSTM)-based denoiser for illustrative purpose. As illustrated in Fig. 1, the denoising network contains two combined layers from bottom-up followed by a mask estimation layer. Each combined layer consists of a BLSTM layer and a Rectifier Linear Unit (ReLU) activation function layer. Applying magnitude of the noisy speech (denoted by $|\mathbf{Z}|$) as the input feature, the network is trained to predict the phase sensitive mask [24] for the target speech (denoted by $\hat{\mathbf{M}}$) via the magnitude and temporal spectrum approximation loss, defined by [25]:

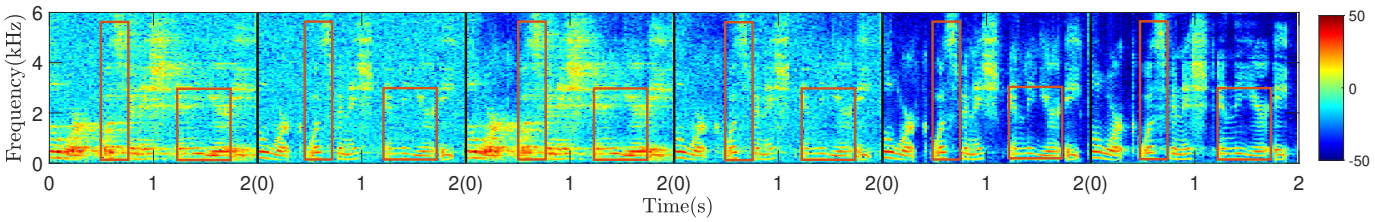
$$\begin{aligned}\mathcal{J}_{\mathrm{Denoiser}} =& \frac{1}{N}\sum\Big(\parallel \hat{\mathbf{M}} \odot |\mathbf{Z}| - |\mathbf{A}| \odot \cos(\theta_{\mathbf{Z}} - \theta_{\mathbf{A}}) \parallel_F^2 \\ &+ w_d \parallel f_d(\hat{\mathbf{M}} \odot |\mathbf{Z}|) - f_d(|\mathbf{A}| \odot \cos(\theta_{\mathbf{Z}} - \theta_{\mathbf{A}})) \parallel_F^2 \\ &+ w_c \parallel f_c(\hat{\mathbf{M}} \odot |\mathbf{Z}|) - f_c(|\mathbf{A}| \odot \cos(\theta_{\mathbf{Z}} - \theta_{\mathbf{A}})) \parallel_F^2 \Big),\end{aligned} \tag{21}$$

where $\odot$ is the Hadamard product, $|\mathbf{A}|$ is the magnitude of the clean speech, and $\theta_{\mathbf{Z}}$ and $\theta_{\mathbf{A}}$ represent phase angles of the noisy speech and the clean speech respectively. Taking the dynamic information into consideration, we employ the functions (i.e., $f_d(\cdot)$ and $f_c(\cdot)$) to calculate increment and acceleration [26], with the weights w_d and w_c set to 4.5 and 10.0 respectively. The network is independently trained, and can be plugged into the proposed framework.

IV. Experimental Results

In this section, we validate the proposed method and compare it with other methods in several respects.

Data-driven prior construction: To train a blind denoiser, we added the white Gaussian noise to the clean speech signals



(a) Observed speech (b) WPE (c) Denoiser only (d) Denoiser+WPE (e) Proposed (f) Ground truth

Fig. 2. The visualization results of all comparison methods at SNR = 20 dB and T_{60} = 786 ms, with the speech truncated to 2 s.

TABLE I
THE RESULTS OF ALL COMPARISON METHODS UNDER THE SCENARIOS OF REVERBERATION WITH NOISE OF DIFFERENT LEVELS. THE BEST RESULTS ARE IN BOLD AND THE SECOND BEST RESULTS ARE UNDERLINED.

SNR (dB)	Methods	SDR (dB)	STOI	PESQ	CD
0	Observed	-3.87	0.46	1.33	8.31
	WPE	-4.33	0.49	1.29	8.42
	Denoiser only	0.94	0.51	1.95	7.71
	Denoiser+WPE	3.65	0.52	**2.03**	**6.67**
	Proposed ($\rho = 12.08$, $\mu = 0.15$)	**5.58**	**0.58**	1.80	7.03
10	Observed speech	0.35	0.59	1.81	7.61
	WPE	3.76	0.65	1.80	7.90
	Denoiser only	1.29	0.66	2.00	7.30
	Denoiser + WPE	8.67	0.70	**2.32**	**5.06**
	Proposed ($\rho = 10.30$, $\mu = 0.55$)	**10.06**	**0.74**	2.11	6.91
20	Observed speech	1.14	0.66	1.97	6.13
	WPE	11.74	0.79	2.40	6.85
	Denoiser only	1.20	0.68	1.97	4.64
	Denoiser + WPE	13.02	0.81	2.51	4.76
	Proposed ($\rho = 15.25$, $\mu = 0.55$)	**14.78**	**0.84**	**2.91**	**4.21**
30	Observed speech	1.23	0.69	2.02	4.81
	WPE	15.64	0.86	2.81	5.36
	Denoiser only	1.24	0.70	2.02	4.32
	Denoiser + WPE	15.64	0.86	2.84	4.42
	Proposed ($\rho = 7.81$, $\mu = 0.13$)	**16.45**	**0.87**	**3.19**	**3.81**
40	Observed speech	1.24	0.70	2.04	4.27
	WPE	16.68	**0.88**	3.06	4.03
	Denoiser only	1.24	0.70	2.02	4.45
	Denoiser + WPE	16.64	**0.88**	3.09	3.55
	Proposed ($\rho = 6.32$, $\mu = 0.29$)	**17.03**	**0.88**	**3.21**	**3.37**

randomly chosen from the Wall Street Journal dataset [27]. A training set with 20,000 utterances and a validation set with 5000 utterances at various signal-noise ratios (SNRs) between -5 dB and 40 dB were obtained. Each utterance was split into 4 s with a sampling rate of 16 kHz. We implemented the BLSTM-based denoiser based on Adam optimizer [28] with an initial learning rate of 0.0005 and a mini-batch of 32 to minimize the loss function (21) in 60 epochs. The proposed framework was implemented in the STFT domain using a Hann window, where the frame length was 32 ms of 75% overlapping. For the proposed framework, we set $L = 16$, $D = 2$ and $I = 5$ respectively in our experiments.

Method comparison and evaluation: To test the method, we generated a test set by randomly choosing 4-channel speech

TABLE II
THE COMPARISON RESULTS OF WPE AND THE PROPOSED METHOD UNDER THE SCENARIO OF REVERBERATION WITHOUT NOISE.

T_{60} (ms)	Methods	SDR (dB)	STOI	PESQ	CD
265	Observed	14.66	0.93	2.97	2.14
	WPE	15.54	**0.95**	3.24	2.20
	Proposed	**15.62**	**0.95**	**3.40**	**2.05**
419	Observed	7.18	0.74	2.08	4.16
	WPE	18.23	**0.94**	3.18	3.27
	Proposed	**18.31**	**0.94**	**3.23**	**3.13**
786	Observed	1.24	0.71	2.04	4.25
	WPE	16.87	**0.90**	**3.19**	**2.98**
	Proposed	**17.16**	**0.90**	**3.19**	2.99

from Libri-adhoc40-simu corpus [29] with the reverberation time (T_{60}) being 265 ms, 419 ms and 786 ms. Both noise-free and noisy cases were tested. Considering the scenario of reverberation with noise, we added the white Gaussian noise to the convolved speech of T_{60} = 786 ms, and SNRs were set to 0 dB, 10 dB, 20 dB, 30 dB, and 40 dB respectively. We compared the vanilla WPE, denoiser-only method, intuitive concatenation of the denoiser and WPE (denoted by `Denoiser+WPE`) and the proposed method.

For evaluation metrics, we adopted signal to distortion ratio (SDR) [30], perceptual evaluation of speech quality (PESQ) [31], short-time objective intelligibility (STOI) [32] and cepstral distance (CD) [33] in our experiments. In general, for SDR, PESQ and STOI, larger values indicate better performance, while for CD, smaller values indicate better performance.

Results: The comparison results of all evaluation metrics in the scenario of reverberation with additive white Gaussian noise are reported in Table I. From the table, we can see that the proposed method outperforms other compared methods in most noisy cases. Although the PESQ and CD values of the proposed method are relatively inferior to those of `Denoiser+WPE` at SNR = 0 dB and SNR = 10 dB, the SDR and STOI values of the former perform better than the latter. For visual comparison, we take the scenario SNR = 20 dB as an example, shown in Fig. 2.

We further compare the performance of the vanilla WPE with the proposed method in the scenario of reverberation without noise. Table II lists the results, where ρ and μ are set to 7.5 and 0.08 for the proposed method respectively.

From the overall comparison results, we can conclude that our proposed algorithm still maintains its advantages, with slightly better performance compared to the vanilla WPE in purely reverberant conditions.

V. Conclusion

In this paper, we proposed a method for incorporating data-driven speech priors to improve the performance and robustness of WPE. For this purpose, a PnP strategy based on variable splitting with ADMM, specifically the RED strategy, was employed. A BLSTM-based denoiser was designed and plugged into the optimization steps for capturing the prior from data. We found experimentally that the proposed method can effectively handle reverberation scenarios with or without additive noise. Future research investigates in depth the effects of various denoisers and the performance of the method under non-Gaussian noise.

References

[1] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 1492–1501, Jul. 2017.

[2] O. Schwartz, S. Gannot, and E. A. Habets, "An expectation-maximization algorithm for multimicrophone speech dereverberation and noise reduction with coherence matrix estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, pp. 1495–1510, Sep. 2016.

[3] S. Inoue, H. Kameoka, L. Li, S. Seki, and S. Makino, "Joint separation and dereverberation of reverberant mixtures with multichannel variational autoencoder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 96–100, 2019.

[4] D. Schmid, G. Enzner, S. Malik, D. Kolossa, and R. Martin, "Variational bayesian inference for multichannel dereverberation and noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, pp. 1320–1335, Aug. 2014.

[5] K. Kinoshita *et al.*, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *in Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust.*, pp. 1–4, IEEE, 2013.

[6] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multi-channel equalization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, pp. 680–693, Apr. 2016.

[7] S. Braun *et al.*, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, pp. 1056–1071, Jun. 2018.

[8] G. Huang, J. Benesty, I. Cohen, and J. Chen, "A simple theory and new method of differential beamforming with uniform linear microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1079–1093, Mar. 2020.

[9] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 85–88, 2008.

[10] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 18, pp. 1717–1731, Sep. 2010.

[11] A. Jukić, T. van W., T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 1509–1520, Jun. 2015.

[12] M. Witkowski and K. Kowalczyk, "Split bregman approach to linear prediction based dereverberation with enforced speech sparsity," *IEEE Signal Process. Lett.*, vol. 28, pp. 942–946, Apr. 2021.

[13] K. Han *et al.*, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 982–992, Mar. 2015.

[14] Z. Wang and D. Wang, "Deep learning based target cancellation for speech dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 941–950, Feb. 2020.

[15] N. Shlezinger, Y. C. Eldar, and S. P. Boyd, "Model-based deep learning: On the intersection of deep learning and optimization," *IEEE Access*, vol. 10, pp. 115384–115398, Nov. 2022.

[16] Y. Bai, W. Chen, J. Chen, and S. Guo, "Deep learning methods for solving linear inverse problems: Research directions and paradigms," *Signal Process.*, vol. 177, p. 107729 (23 pages), 2020.

[17] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Process. Mag.*, pp. 18–44, 2021.

[18] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online wpe dereverberation.," in *Interspeech*, pp. 384–388, Aug. 2017.

[19] J. Chen, M. Zhao, X. Wang, C. Richard, and S. Rahardja, "Integration of physics-based and data-driven models for hyperspectral image unmixing," *IEEE Signal Process. Mag.*, to appear.

[20] X. Wang, J. Chen, Q. Wei, and C. Richard, "Hyperspectral image super-resolution via deep prior regularization with parameter estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1708–1723, 2021.

[21] S. H. Chan, X. Wang, and O. A. Elgendy, "Plug-and-play ADMM for image restoration: Fixed-point convergence and applications," *IEEE Trans. Comput. Imaging*, vol. 3, no. 1, pp. 84–98, 2016.

[22] Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising (RED)," *SIAM J. Imaging Sci.*, vol. 10, no. 4, pp. 1804–1844, 2017.

[23] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, pp. 1–122, Jan. 2011.

[24] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, pp. 1702–1726, May. 2018.

[25] C. Xu, W. Rao, E. S. Chng, and H. Li, "Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 6990–6994, 2019.

[26] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 34, pp. 52–59, Feb. 1986.

[27] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete ldc93s6a," *Web Download. Philadelphia: Linguistic Data Consortium*, vol. 83, 1993.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] S. Guan *et al.*, "Libri-adhoc40: A dataset collected from synchronized ad-hoc microphone arrays," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1116–1120, 2021.

[30] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 1462–1469, Jun. 2006.

[31] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, pp. 749–752, 2001.

[32] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 2125–2136, Feb. 2011.

[33] K. Kinoshita *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, pp. 1–19, 2016.