Speech Dereverberation with Deconvolution Regularized by Denoising

Haonan Hu, Ziye Yang, Jie Chen, Lijun Zhang

Research and Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen, China
CIAIC, School of Marine Science and Technology, Northwestern Polytechnical University, China
{hnhu,yzy97}@mail.nwpu.edu.cn, dr.jie.chen@ieee.org, zhanglj7385@nwpu.edu.cn

***Abstract*—Deconvolution-based speech dereverberation continues to present challenges due to the difficulties in accurately acquiring Room Impulse Responses (RIRs) and the inherently ill-conditioned nature of deconvolution. Despite advancements in RIR measurement and estimation, substantial room for improvement remains in addressing the latter challenge. This paper proposes a novel prior-driven dereverberation framework utilizing Regularization by Denoising (RED) to incorporate data priors into the deconvolution process, thereby addressing this persistent challenge. Specifically, we formulate the dereverberation process via an optimization problem with the additional regularizer and the Half Quadratic Splitting (HQS) strategy is then utilized to solve the optimization problem. Experimental validation conducted on both the RIR simulation platform pyroomacoustics and the realistic acoustics platform SoundSpaces demonstrates the efficacy of our framework, even in the presence of environmental noise and RIR errors.**

***Index Terms*—Speech dereverberation, ill-conditioned deconvolution, deep priors, Regularization by Denoising, SoundSpaces**

I. INTRODUCTION

In enclosed spaces, speech signals propagate with inevitable energy attenuation and are subject to reflections off surfaces and objects, resulting in an acoustic phenomenon known as reverberation. This phenomenon can significantly degrade speech quality and intelligibility. Moreover, reverberation may adversely impact the performance of Automatic Speech Recognition (ASR) systems, particularly in scenarios compounded by additive noise. As a result, speech dereverberation technology has garnered considerable interest within the speech signal processing community.

Reverberation is physically modeled by convolving an anechoic speech utterance with an acoustic path, naturally leading to the use of deconvolution strategies. However, speech dereverberation has historically faced two primary challenges. Firstly, in real-world settings, the accurate acquisition and estimation of room impulse responses (RIRs), which is crucial for analyzing acoustic environments, present significant challenges [1], [2]. Secondly, even with available RIR data, the deconvolution operation for speech dereverberation remains an ill-conditioned inverse problem [3], [4], rendering inverse processing difficult, if not impossible. This difficulty is particularly pronounced in the presence of unavoidable environmental noise and minor RIR perturbations, which can significantly contribute to erroneous results. Recent advancements in the utilization of data-driven and statistical methods for RIR measurement and estimation [5]–[7] have contributed to progress in addressing the first challenge. However, there has been limited, if any, effort directed towards effectively tackling the second challenge—the ill-conditioned problem associated with speech deconvolution.

In fact, the ill-conditioned deconvolution problem encountered in speech dereverberation bears similarities to those encountered in image deconvolution/deblurring [8], which have been successfully addressed through the significant utilization of image priors. Among candidate strategies, the plug-and play method [9]–[15] shows its superior effectiveness in image and speech processing. This approach aims to incorporate data priors into the optimization iterations via a deep denoising algorithm, thereby facilitating the solution process of ill-conditioned deconvolution problems.

Inspired by this advance, our work aims to introduce a prior-driven deconvolution-based dereverberation framework to more effectively address the ill-conditioned problem under reverberant-noisy scenarios. Specifically, we formulate the speech dereverberation process through deconvolution using an optimization problem integrating an additional regularizer. This regularizer is not explicitly handcrafted; instead, we employ the Plug-and-Play (PnP) strategy, specifically the Regularization by Denoising (RED) strategy [16], to incorporate deep priors extracted from data. Additionally, we introduce the Half Quadratic Splitting (HQS) method to further solve the optimization problem. Extensive experiments on the widely employed RIR simulation platform, pyroomacoustics[17], and the highly realistic acoustics platform, SoundSpaces [18], [19], to evaluate the performance of the proposed framework.

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. The operator $(\cdot)^\top$ denote matrix transpose and $(\cdot)^*$ denote conjugate, and $\langle\cdot\rangle$ represents the inner product.

II. PROBLEM FORMULATION

We consider the scenario where a single-channel microphone receives the reverberant speech interfered by the additive noise. The captured signal $y(t)$ can be modeled by the

The work was supported in part by Shenzhen Science and Technology Program JCYJ20230807145600001, TCL science and technology innovation fund, Shaanxi Key Industrial Innovation Chain Project 2022ZDLGY01-02.

following equation:

$$y(t) = h(t) * s(t) + n(t) \tag{1}$$

where t indexes discrete time and $s(t)$ represents the clean speech signal. $h(t)$ represents the time-invariant RIR, characterizing all reflections and attenuations along the propagation path from the source to the microphone. $n(t)$ represents the additive noise, and $*$ denotes the convolution operation. To facilitate the method presentation, (1) is often written in the following form as

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n}, \tag{2}$$

where $\mathbf{s}$ and $\mathbf{y}$ denotes the vector forms for clean speech and observed signals, and $\mathbf{H}$ is the convolutive matrix formed according to $h(t)$. Then the fidelity term of dereverberation process can be formulated by an inverse problem such that

$$\min_{\mathbf{s}} \|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2. \tag{3}$$

Given (3) is inherently ill-conditioned, it is essential to incorporate the regularization term to stabilize the solution process and enhance the quality of the solution of this inverse problem, which can be written as

$$\min_{\mathbf{s}} \|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2 + \mathcal{R}(\mathbf{s}) \tag{4}$$

where $\mathcal{R}(\mathbf{s})$ represents the regularization term. Recently, considerable effort has been invested in meticulously designing the appropriate regularization term so as to integrate priors into the optimization problem, such as the sparsity prior [20] and the Complex Generalized Gaussian Prior [21]. Nevertheless, it is non-trivial to handcraft such effective regularizer for (4) with the efficient solution methodology.

III. THE PROPOSED FRAMEWORK

In this work, rather than elaborately handcrafting the regularizer, we propose to directly derive priors from speech data to facilitate the problem in (4). Specifically, we consider the regularization in the form as

$$\mathcal{R}(\mathbf{s}) = \frac{1}{2}\langle \mathbf{s}, \mathbf{s} - f(\mathbf{s})\rangle, \tag{5}$$

where $f(\cdot)$ denotes an off-the-shelf denoiser. This form, called RED, measures the inner product between the desired speech and its denoised residual [16]. Different from the prototype PnP strategy, RED leverages a specific formulation and demonstrates advantageous derivative characteristics under mild assumptions. Integrating RED into (4), we can obtain the problem formulated as

$$\min_{\mathbf{s}} \|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2 + \frac{\rho}{2}\mathbf{s}^\top(\mathbf{s} - f(\mathbf{s})) \tag{6}$$

with ρ being the regularization parameter.

To solve the problem in (6), we initially employ the variable splitting technique to introduce an auxiliary variable $\mathbf{z}$, leading to the following equivalent constrained optimization formulation

$$\begin{aligned} &\min_{\mathbf{s}} \|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2 + \frac{\rho}{2}\mathbf{z}^\top(\mathbf{z} - f(\mathbf{z})) \\ &\text{s.t.} \quad \mathbf{z} = \mathbf{s}. \end{aligned} \tag{7}$$

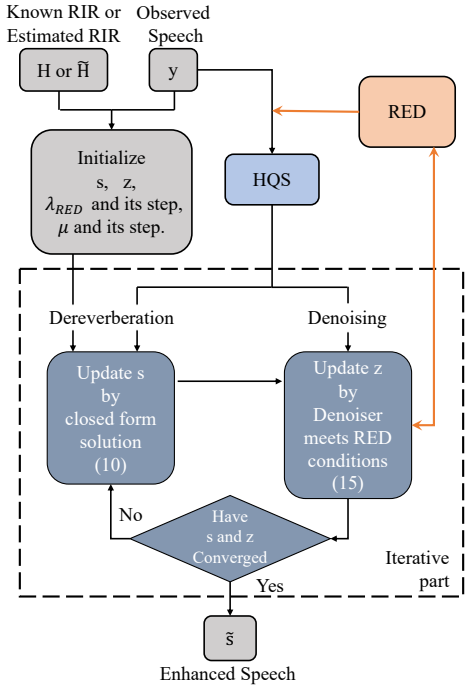

Fig. 1. Block diagram of the proposed framework.

Subsequently, we tackle (7) via the HQS algorithm [22], involving an additional quadratic penalty term

$$\mathcal{L}(\mathbf{s}, \mathbf{z}) = \|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2 + \frac{\lambda}{2}\|\mathbf{s} - \mathbf{z}\|^2 + \frac{\rho}{2}\mathbf{z}^\top(\mathbf{z} - f(\mathbf{z})), \tag{8}$$

where λ denotes the penalty parameter. The variables in (8) can then be optimized through the following alternating minimization problems with respect to $\mathbf{x}$ and $\mathbf{z}$ respectively, with $(\cdot)^{(n)}$ in the following content denoting the nth iteration.

1) Optimization with respect to $\mathbf{s}$*:* The optimization of (8) becomes:

$$\mathbf{s}^{(n+1)} = \arg\min_{\mathbf{s}} \|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2 + \frac{\lambda}{2}\|\mathbf{s} - \mathbf{z}^{(n)}\|^2. \tag{9}$$

Given the assumption that $\mathbf{H}$ is available, the closed-form solution to (9) regarding $\mathbf{s}$ can be derived by

$$\mathbf{s}^{(n+1)} = \left(\mathbf{H}^\top\mathbf{H} + \frac{\lambda}{2}I\right)^{-1}\left(\mathbf{H}^\top\mathbf{y} + \frac{\lambda}{2}\mathbf{z}^{(\mathbf{n})}\right). \tag{10}$$

2) Optimization with respect to $\mathbf{z}$*:* The optimization problem (8) reduces to:

$$\mathbf{z}^{(n+1)} = \arg\min_{\mathbf{z}} \frac{\lambda}{2}\left\|\mathbf{s}^{(n+1)} - \mathbf{z}\right\|^2 + \frac{\rho}{2}\mathbf{z}^\top(\mathbf{z} - f(\mathbf{z})) \tag{11}$$

Setting the derivative of the above optimization w.r.t. $\mathbf{z}$ to zero, it becomes feasible to directly minimize it through an iterative scheme, leading to the following equation:

$$\lambda(\mathbf{z}^{(n+1)} - \mathbf{s}^{(n+1)}) + \rho(\mathbf{z}^{(n+1)} - f(\mathbf{z}^{(n+1)})) = 0. \tag{12}$$

The solution to this problem can be achieved via the fixed-point iteration:

$$\lambda(\mathbf{z}^{(n+1,i)} - \mathbf{s}^{(n+1)}) + \rho(\mathbf{z}^{(n+1,i)} - f(\mathbf{z}^{(n+1,i)})) = 0, \tag{13}$$

leading to

$$\mathbf{z}^{(n+1,i)} = \frac{\lambda}{\lambda+\rho}\mathbf{s}^{(n+1,i)} + \frac{\rho}{\lambda+\rho} f(\mathbf{z}^{(n+1,i)}). \tag{14}$$

For simplification, this can be rewritten as:

$$\mathbf{z}^{(n+1,i)} = \mu\mathbf{s}^{(n+1,i)} + (1-\mu) f(\mathbf{z}^{(n+1,i)}), \tag{15}$$

with $\mu = \frac{\lambda}{\lambda+\rho}$ and the inner iteration $i = 1, \cdots, I$, where μ serves as a balance parameter between the estimate provided by $\mathbf{s}^{(n+1)}$ and the denoiser $f(\mathbf{z})$.

The overall algorithm is illustrated in Fig. 1, where variables $\mathbf{s}$ and $\mathbf{z}$ are updated iteratively until convergence.

Remark 1: Though mathematically easy to follow, $\mathbf{H}^\top\mathbf{H}$ in matrix form results in inefficient computational processes. Computing in frequency domain presents to be more effective. In the frequency domain, the signal model in (1) writes

$$Y(f) = H(f)S(f) + N(f) \tag{16}$$

with f indicating the frequency bin, where $Y(f)$, $H(f)$, $S(f)$ and $N(f)$ are the counterparts of $y(t)$, $h(t)$, $s(t)$ and $n(t)$ in the frequency domain respectively. This way, the solution w.r.t. $S(f)$ in the frequency domain is given by:

$$S^{(n+1)}(f) = \left(|H(f)|^2 + \frac{\lambda}{2}\right)^{-1}\left(H^*(f)Y(f) + \frac{\lambda}{2}Z^{(n)}(f)\right), \tag{17}$$

which effectively reduces the computational burden.

Remark 2: Within the proposed framework, empirical evidence suggests that the adjustment of dynamic parameters can enhance the overall performance. Specifically, we implement dynamical values λ and μ, which are progressively increased from an initial value with the predefined step size.

IV. Experiments

In this section, we first introduce the simulation environments, followed by the details on the experimental settings, and finally discuss the results in several respects.

A. Experimental environments

Simulated scenarios based on pyroomacoustics: In the experiment, we simulated a scenario within a room configured to the dimension of 5 meters (m) × 4 m × 6 m, where a single microphone is placed at a randomly chosen position. Clean utterances, randomly selected from the Voice Bank corpus [23], were convolved with the RIRs generated via pyroomacoustics [17] to synthesize reverberant speech, where the reverberation times (T_{60}) were set to 190 millisecond (ms), 430 ms and 890 ms respectively. To test the method under varying levels of the additive noise, we considered white Gaussian noise (WGN) and added it to the reverberant speech with the signal-to-ratios (SNRs) setting to 0 dB, 10 dB, and 20 dB respectively.

Moreover, it is inevitable that various types of errors will manifest in the estimation of RIR, it is thus necessary to assess the efficacy of the proposed method under such conditions of uncertainty. Specifically, to simulate the scenario of RIR estimation errors, the accurate RIR ($h(t)$) generated by pyroomacoustics was replaced with an approximate representation, denoted as $\tilde{h}(t)$, given by

$$\tilde{h}(t) = h(t) + (\alpha|h(t)|)\, z(t). \tag{18}$$

The term $(\alpha|h(t)|)z(t)$ introduces a magnitude related perturbation to $h(t)$, with α a scaling constant $z(t)$ a random error [24].

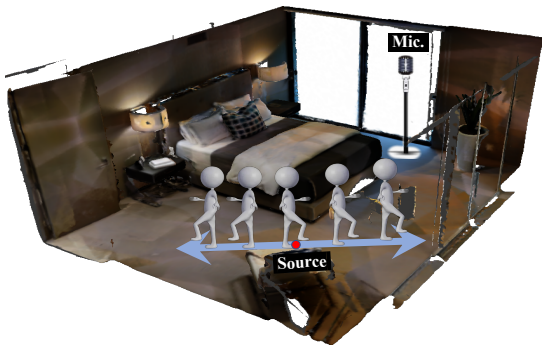

Fig. 2. An example of the experimental scenes simulated on SoundSpaces, showing positions of source and microphone. The red dot indicates the reference position of the sound source, and the source moves in the directions of the bidirectional arrow.

Simulated scenarios based on SoundSpaces: To assess the method in real-world scenarios with more realistic RIR error, we applied SoundSpaces [18], [19], a platform developed by Meta, to simulate another experimental environment. Due to the capability of SoundSpaces to closely mimic real world acoustic environments, it could precisely simulate the sound propagation and reflection across diverse environments, including reverberation and attenuation within rooms, thereby providing a practical and highly realistic testing ground. In addition, SoundSpaces enables the insertion of arbitrary sound sources across various publicly available 3D environments. Therefore, we could simulate the close-to-real-world scenario within closed rooms selected from the Matterport3D dataset [25], where the initial 3-dimensional coordinates of the microphone and source were randomly generated to set the reference position, thereby measuring the exact RIRs.

For this scenario, we also considered to establish the condition of RIR estimation error: First, we set a reference point of source (as mentioned before), serving as the focal point for all source movements. Then, we move the source along a random direction aligned with a straight line extending in both directions of the bidirectional arrow (as indicated by bidirectional arrow in Fig. 2). At each new position reached with a step size of 0.1 m, an RIR measurement was conducted. Afterwards, considering the scenario of reverberation with noise, clean speech signals were convolved with these measured RIRs, followed by the addition of WGN at SNR of 20 dB. During the evaluation, We employed the RIR corresponding to the focal point as $\tilde{\mathbf{H}}$, at each relocated position to derive (10) within the proposed framework.

TABLE I

THE RESULTS OF ALL COMPARISON METHODS IN THE SIMULATED SCENARIO AT DIFFERENT SNRS WITH VARIOUS T_{60}. THE TERM OF ITERATION (ITER.) DENOTES THE QUANTITY OF ITERATIONS REQUIRED FOR THE METHOD TO ACHIEVE CONVERGENCE. THE BEST RESULTS ARE IN BOLD.

		T_{60} →	190 ms				430 ms				890 ms			
SNR (dB)	Error(α) (%)	Methods↓	STOI ↑	PESQ ↑	F-SNR (dB)↑	Iter. ↓	STOI ↑	PESQ ↑	F-SNR (dB)↑	Iter. ↓	STOI ↑	PESQ ↑	F-SNR (dB)↑	Iter. ↓
0 dB	0%	Observed	0.631	1.035	3.772	-	0.548	1.028	3.482	-	0.554	1.048	3.480	-
		JE	0.752	1.074	5.965	81	0.634	1.042	3.400	250	0.728	1.099	6.052	222
		Proposed-s	0.763	1.099	6.318	84	0.648	1.039	3.996	238	0.755	1.156	7.368	270
		Proposed-d	**0.821**	**1.291**	**9.737**	**35**	**0.769**	**1.190**	**7.349**	**38**	**0.784**	**1.332**	**8.782**	**40**
10 dB	0%	Observed	0.718	1.078	5.721	-	0.607	1.034	4.911	-	0.611	1.100	4.496	-
		JE	0.862	1.625	10.636	**36**	0.801	1.154	7.250	105	0.827	1.698	10.796	124
		Proposed-s	0.868	1.683	11.065	**36**	0.811	1.208	8.061	112	0.828	1.721	11.869	103
		Proposed-d	**0.881**	**1.868**	**11.896**	**36**	**0.836**	**1.368**	**10.116**	**32**	**0.848**	**1.761**	**11.994**	**36**
20 dB	0%	Observed	0.772	1.313	7.263	-	0.659	1.095	6.205	-	0.635	1.209	5.149	-
		JE-0	0.897	2.104	11.392	**9**	0.879	1.902	12.321	40	0.873	2.109	13.483	**5**
		Proposed-s	**0.901**	2.206	11.430	10	**0.888**	2.049	12.932	**36**	0.881	2.482	14.962	24
		Proposed-d	0.900	**2.266**	**11.464**	10	**0.888**	**2.118**	**13.009**	**36**	**0.883**	**2.644**	**15.489**	8
	15%	JE	0.891	2.033	**11.050**	**7**	0.882	1.774	11.050	44	0.862	1.887	12.010	43
		Proposed-d	**0.896**	**2.188**	11.004	9	**0.888**	**2.068**	**11.759**	**36**	**0.876**	**2.229**	**12.455**	**8**

B. Experimental settings

Deep prior construction: Benefiting from the flexibility of the proposed framework, it permits the integration of any denoiser to incorporate deep speech priors. Therefore, we directly utilized a pre-trained denoiser, known as speech enhancement generative adversarial network (SEGAN) [26], to focus on the scope of this work. This denoiser, which is developed in an end-to-end manner within an adversarial framework, has been trained on an extensive noisy speech dataset, thereby deemed effective for speech denoising in complex noisy scenarios.

Method comparison and evaluation: For the method comparison, here we consider the joint enhancement framework [24] (denoted by `JE`), which utilizes the PnP strategy to simultaneously execute dereverberation and denoising within a unified framework. Furthermore, within the proposed method, different strategies for parameter selection are evaluated, including the static strategy (denoted by `Proposed-s`) and the dynamic strategy (denoted by `Proposed-d`).

To quantify the experimental results, we utilize the objective metrics including short-time objective intelligibility measure (STOI)[27], perceptual evaluation of speech quality (PESQ)[28] and frequency weighted segmental SNR (F-SNR)[29], to evaluate speech intelligibility and quality. Beside, the Word Error Rate (WER) is also employed to verify the influence of the proposed method on the backend ASR system, which was facilitated through the SpeechRecognition library[30].

Implementation details: Concerning the parameter settings within the framework, λ is set to 2.2 and μ is set to 0.28. For `Proposed-s`, the parameter settings remain static throughout the iterations. Conversely, for `Proposed-d`, these settings gradually increase from their respective base values with the step sizes of 0.28 for λ and 0.015 for μ. For all the above scenarios, the test speech signals are sampled at 16 kHz. Specifically, to simulate the scenario of RIR estimation errors, the accurate RIR $h(t)$ generated by pyroomacoustics is replaced with an approximate representation $\tilde{h}(t)$, perturbated by (18), where the scaling constant α is set from 0 to 0.5 (0% to 50%), and the random error $z(t)$ is modeled using a Gaussian distribution with zero mean and unit variance $z(t) \sim \mathcal{N}(0, 1)$. When performing the calculation of the "Optimization with respect to s" step in the frequency domain for more efficient computing, we utilize Fast Fourier Transform (FFT) with the following parameter settings: To capture the short-time characteristics of the speech signal, the frame length is set to 25 ms (400 samples at a sampling rate of 16 kHz) with a frame shift of 10 ms (160 samples) with Hamming window applied. The FFT length is chosen to be 512, which is a power of two and greater than the frame length, ensuring efficient computation. To ensure smooth transitions between frames, a 50% overlap is used, meaning each frame overlaps with the previous frame by 200 samples.

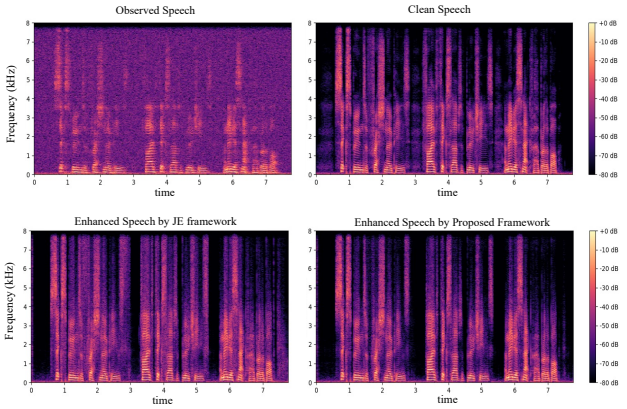

Fig. 3. Spectograms of observed (top left), clean (top right), `JE` (Bottom Left) and the proposed method (bottom right) in the simulated scenario.

C. Results discussion

From Table I, we find that the proposed method surpasses `JE` across all the scenarios, particularly under the condition

TABLE II
THE WER RESULTS OF PROPOSED-D AND JE IN THE SIMULATED SCENARIO AT SNR=20 DB.

	$T_{60}\rightarrow$	190 ms	430 ms	890 ms
Error (%)	Methods↓	WER (%) ↓		
0%	JE	27	25	41
25%	JE	30	29	48
0%	Proposed-d	24	20	32
25%	Proposed-d	24	25	32

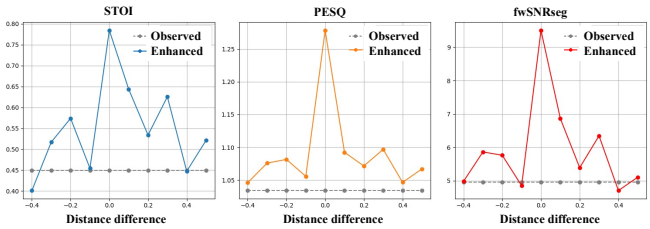

Fig. 4. Trend of F-SNR (left), PESQ (middle) and STOI (right) considering the RIR measurement errors on the SoundSpaces Platform.

characterized by high T_{60} and intense WGN. Additionally, the efficacy of the dynamic parameter adjustment strategy is validated, as `Proposed-d` yields almost all the best results. Meanwhile, it is evident from the values of `Iter.` that under a consistent convergence criterion, the proposed framework converges more rapidly than the comparison methods. Moreover, for visual comparison, we take the scenario of SNR = 20 dB and T_{60} = 430 ms as an example, and illustrate the spectrograms in Fig. 3. From the spectrograms, we can conclude that the inclusion of the RED strategy provides valuable priors for the deconvolution task, which helps to better preserve the naturalness and clarity of the speech signal during the dereverberation process.

In order to explore the impact of dereverberation methods on the performance of back-end ASR, we in addition present the WER results of `JE` and `Proposed-d` in Table II. From the findings, it is apparent that the integration of `Proposed-d` exhibits a notable enhancement in the efficacy of the ASR system, regardless of the presence or absence of errors in RIR estimation.

Fig. 4 presents the results under the scenarios simulated based on SoundSpaces. From these results, we can see that if the RIR measured at alternative locations is considered as the RIR estimation at the current location, which corresponds to the scenarios of RIR estimation with errors, our proposed framework still exhibits a capacity for dereverberation to a certain extent.

V. CONCLUSION

In this paper, we proposed a prior-driven dereverberation framework, providing an effective solution to the ill-conditioned deconvolution problem. Experiments were conducted on two respective platforms, including SoundSpaces, a realistic acoustic simulation platform. The results demonstrate that our framework is capable of solving ill-conditioned problem and boosting the speech dereverberation performance, even in the presence of environmental noise and RIR errors. Future work will focus on updating RIR estimates during iterations to further improve the performance.

REFERENCES

[1] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *Journal of the Audio engineering society*, vol. 50, pp. 249–262, 2002.

[2] N. Hahn and S. Spors, "Comparison of continuous measurement techniques for spatial room impulse responses," *2016 24th European Signal Processing Conference (EUSIPCO)*, pp. 1638–1642, 2016.

[3] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *The Journal of the Acoustical Society of America*, vol. 66, pp. 165–169, Jul. 1979.

[4] J. Mourjopoulos, "On the variation and invertibility of room impulse response functions," *Journal of Sound and Vibration*, vol. 102, pp. 217–228, 1985.

[5] N. Mohanan, R. Velmurugan, and P. Rao, "Speech dereverberation using nmf with regularized room impulse response," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4955–4959, 2017.

[6] C. J. Steinmetz, V. K. Ithapu, and P. Calamia, "Filtered noise shaping for time domain room impulse response estimation from reverberant speech," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 221–225, 2021.

[7] M. Pezzoli, D. Perini, A. Bernardini, F. Borra, F. Antonacci, and A. Sarti, "Deep prior approach for room impulse response reconstruction," *Sensors*, vol. 22, p. 2710, 2022.

[8] X. Wang, J. Chen, and C. Richard, "Tuning-free plug-and-play hyperspectral image deconvolution with deep priors," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[9] J. Chen, M. Zhao, X. Wang, C. Richard, and S. Rahardja, "Integration of physics-based and data-driven models for hyperspectral image unmixing: A summary of current methods," *IEEE Signal Processing Magazine*, vol. 40, pp. 61–74, Mar. 2023.

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

[11] Z. Yang, W. Yang, K. Xie, and J. Chen, "Integrating data priors to weighted prediction error for speech dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3908–3923, 2024.

[12] Z. Yang, W. Yang, K. Xie, and J. Chen, "Speech dereverberation using weighted prediction error with prior learnt from data," *European Signal Processing Conference (EUSIPCO)*, pp. 356–360, 2023.

[13] Z. Yang, J. Chen, C. Richard, and J. Li, "Plug-and-play wpe guided by deep spectrum estimation for speech dereverberation," *IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, 2024.

[14] C. Chang, Z. Yang, and J. Chen, "Plug-and-play mvdr beamforming for speech separation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1346–1350, 2024.

[15] K. Matsumoto and K. Yatabe, "Determined bss by combination of iva and dnn via proximal average," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[16] Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising (red)," *SIAM Journal on Imaging Sciences*, vol. 10, pp. 1804–1844, 2017.

[17] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 351–355, 2018.

[18] C. Chen, U. Jain, C. Schissler, *et al.*, "Soundspaces: Audio-visual navigation in 3d environments," *Computer Vision–ECCV*, pp. 17–36, Aug. 2020.

[19] C. Chen, C. Schissler, S. Garg, *et al.*, "Soundspaces 2.0: A simulation platform for visual-acoustic learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8896–8911, 2022.

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

[21] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 1509–1520, 2015.

[22] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, pp. 1–122, 2011.

[23] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," *International Conference Oriental COCOSDA held jointly with Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, Nov. 2013. DOI: 10.1109/icsda.2013.6709856.

[24] A. Raikar, S. Basu, L. Pandey, and R. Hegde, "Multi-channel joint dereverberation and denoising using deep priors," *IEEE India Council International Conference (INDICON)*, pp. 1–6, 2018.

[25] A. Chang, A. Dai, T. Funkhouser, *et al.*, "Matterport3d: Learning from rgb-d data in indoor environments," *2017 International Conference on 3D Vision (3DV)*, pp. 667–676, 2017.

[26] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[27] 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*, pp. 2125–2136, Sep. 2011. DOI: 10.1109/tasl.2011.2114881.

[28] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, pp. 749–752, 2001.

[29] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, pp. 1462–1469, Jul. 2006. DOI: 10.1109/tsa.2005.858005.

[30] A. Zhang, "Speechrecognition: Speech recognition module for python, supporting several engines and apis, online and offline," *GitHub repository*, 2022.