Bifurcation and Reunion: A Loss-Guided Two-Stage Approach for Monaural Speech Dereverberation

Xiaoxue Luo¹,², Chengshi Zheng¹,², Andong Li¹,², Yuxuan Ke¹,², Xiaodong Li¹,²

¹Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
²University of Chinese Academy of Sciences, Beijing, China

{luoxiaoxue, cszheng, liandong, keyuxuan, lxd}@mail.ioa.ac.cn

Abstract

Speech dereverberation is challenging for various speech processing systems. Recently, phase recovery is proved to be significant for improving speech quality and intelligibility, and numerous supervised speech dereverberation algorithms begin focusing on complex spectrum estimation. However, these methods recover clean speech phase at the expense of severe magnitude distortion due to the magnitude-phase compensation effect. To address this problem, we propose a novel loss-guided two-stage framework to progressively guide the process of complex spectrum recovery. In the first stage, a bifurcated network is proposed to separately optimize the magnitude and phase of the complex spectrum coarsely by two distinct loss functions. After that, a reunited network is devised to exploit the complementary characteristics of previous estimations and further refine the complex spectrum. A mathematical derivation is presented to reveal the magnitude-phase compromise phenomenon and validate the rationality of the proposed objective optimization strategy. Experimental results demonstrate that the proposed method improves both speech quality and intelligibility in the dereverberation task, and outperforms other baseline methods.

Index Terms: monaural speech dereverberation, loss-guided, two-stage network, complex domain, phase recovery

1. Introduction

In indoor environments, clean speech is inevitably distorted by reverberation, which severely degrades speech quality and intelligibility [1], affecting the user’s experience of utilizing speech-related devices, such as teleconference systems and hearing aids [2]. To address this problem, numerous traditional dereverberation algorithms have been proposed in the past several decades [3, 4, 5]. More recently, the developments of deep neural networks (DNNs) have greatly improved dereverberation performance by formulating the dereverberation task as a supervised problem, so that the nonlinear mapping relations could be learned from large amounts of training data [6].

For a long time, DNN-based methods mainly focused on estimating the magnitude of the target speech while leaving the phase unaltered [7, 8, 9], which is because the phase shows little regularity and thus makes it difficult to be recovered. Nonetheless, Paliwal et al. [10] pointed out that the recovery of phase information can benefit subjective perception quality, especially under long reverberation time (RT₆₀) conditions. As thus, a handful of complex-domain based single-stage networks have already been proposed in more recent years, where both real and imaginary (RI) parts of the target spectrum are estimated [11, 12, 13]. However, subsequent studies revealed that the RI loss only yields limited phase recovery while causing severe magnitude distortion [14]. To alleviate this problem, an extra magnitude constraint was introduced to the RI loss in [15]. Although such a regularization term can effectively mitigate magnitude distortion by constraining the magnitude optimization space, Wang et al. [16] showed that due to the approximation strategy toward the RI term, the magnitude and phase objectives are entangled in essence and an inaccurate phase estimation may lead to severe magnitude distortion. Therefore, the performance upper bound of single-stage networks is probably limited by the inherent nature of such loss design.

More recently, Li et al. [17, 18] proposed a two-stage network for speech denoising, where spectral magnitude is estimated in the first stage, and the entire complex spectrum is refined afterward. Unlike previous one-stage methods, this two-stage topology explicitly decouples magnitude estimation from the complex spectrum estimation in the intermediate process, which efficiently improves magnitude estimation but still remains limited phase recovery. In this study, different from additive noise, reverberation tends to cause more severe phase corruption due to its convolutive physical property [19]. Therefore, improving the phase estimation accuracy in a step-wise manner is imperative for the speech dereverberation task.

Based on this fact, we propose a novel two-stage speech dereverberation framework in this study, which consists of two subnets, namely bifurcated network (dubbed B-Net) and reunited network (dubbed R-Net). For the first stage, B-Net separately estimates the magnitude and phase of target speech in two branches, obtaining a coarse complex spectrum by coupling the estimated magnitude and phase. Different from the method in [20], two corresponding loss functions are designed for these two branches to tackle magnitude estimation and phase recovery, respectively. For the second stage, R-Net is adopted to refine the estimated coarse complex spectrum. A compromised RI loss with extra magnitude constraint is utilized in this stage to estimate the complex residual details so as to fine-tune the coarsely estimated complex spectrum. The design rationale of our framework lies in three-fold. First, multi-stage topology has shown its advantages in reducing the difficulty of complex spectrum estimation by introducing a sequential optimization strategy [21, 22, 23]. Second, bifurcated topology such as PHASEN [20] is reported efficiently for phase recovery. Third, we systematically give derivations to analyze the trade-off mechanism between magnitude distortion and phase recovery when the RI loss together with magnitude regularization is adopted. It is revealed that the magnitude-phase trade-off can be regulated under the guidance of the well-designed loss

Yuxuan Ke is the corresponding author. This research was funded in part by the National Natural Science Foundation of China (62101550).
functions, which can theoretically guarantee the feasibility of our proposed pipeline.

The remainder of the paper is organized as follows. In Section 2, we describe the proposed system in detail. The experimental setup is presented in Section 3. The results and analysis are given in Section 4. Section 5 concludes this paper.

2. System description

2.1. Problem formulation

The clean speech and the room impulse response (RIR) are \( s[t] \) and \( h[t] \), respectively, and the received reverberated speech signal acquired by a microphone, \( y[t] \), can be modeled as

\[
y[t] = s[t] * h[t] = s[t] * h_d[t] + s[t] * h_r[t],
\]

where \( * \) denotes the convolution operation and \( t \) is the time index. \( h_d[t] \) and \( h_r[t] \) are the direct sound path and reverberation part of \( h[t] \). \( x[t] \) and \( r[t] \) denote the direct speech and its reverberation components, respectively. Short-time Fourier transform (STFT) is adopted to convert the time domain signal into a complex spectrum. After obtaining the estimated IRM, the role of the first stage and that of the second stage are mainly to repair missing details of the coarse complex spectrum. In a nutshell, the whole forward procedure can be formulated as

\[
\begin{align*}
\tilde{X}_{R}^{\text{aux}} & = \mathcal{F}_R(Y_{R}, Y_{I}; \Phi_R), \\
\tilde{\theta}^* & = \arctan(\tilde{X}_{R}^{\text{aux}}, \tilde{X}_{I}^{\text{aux}}), \\
\{X_{R}^{\text{aux}}_1, X_{I}^{\text{aux}}_1\} & = \{X_{R}^{\text{aux}} \cos \tilde{\theta}^*, X_{I}^{\text{aux}} \sin \tilde{\theta}^*\}, \\
\{\tilde{X}_{R}^{\text{aux}}_1, \tilde{X}_{I}^{\text{aux}}_1\} & = \mathcal{F}_B(\tilde{X}_{R}^{\text{aux}}, \tilde{X}_{I}^{\text{aux}}, Y_{R}, Y_{I}; \Phi_B),
\end{align*}
\]

where \( \mathcal{F}_B \) and \( \mathcal{F}_R \) are the mapping functions of the B-Net and the R-Net, respectively. \( \Phi_B \) and \( \Phi_R \) are their corresponding parameter sets.

2.2. System overview

The overall diagram of the proposed system is illustrated in Fig. 1 (a), which consists of two processing stages. In the first stage, we decouple the complex spectrum estimation into two subtasks, i.e., magnitude estimation and phase recovery, by using two decoding branches. Two optimization guidelines are utilized for these two distinct tasks. Note that although magnitude has structural characteristics in the T-F domain, phase shows little regularity, which makes it hard for a network to grasp the distribution if not impossible [20]. Therefore, the phase recovery branch of B-Net estimates an auxiliary complex spectrum, i.e., \( \tilde{X}^{\text{aux}} \), to recover the phase implicitly. The estimated magnitude and phase, i.e., \( |\tilde{X}^{\text{aux}}| \) and \( \tilde{\theta}^* \), in the first stage are then coupled together to re-generate the coarse complex spectrum, i.e., \( \tilde{X}^* \). To further refine the overall complex spectrum, the second stage with R-Net is introduced, receiving both original and coarse complex spectra as input features. Concretely, instead of constructing the spectrum from scratch, a global residual connection is introduced to enforce the network to rely on the residual information.

2.3. B-Net and R-Net

The role of the first stage and that of the second stage are mainly realized through B-Net and R-Net, respectively. Detailed structures of B-Net and R-Net are illustrated in Fig. 1 (b)-(c). Both B-Net and R-Net are similar to the network structure in [12], where the encoder and decoder are utilized to extract spectral features and reconstruct the spectrum, respectively, and a sequence modeling module is inserted in the bottleneck to learn the long-term temporal correlations. For B-Net, the design of such an architecture is inspired by multi-task learning [24, 25], where feature extraction and time sequence modeling are shared across the tasks. Two parallel decoders are devised to separately estimate the ideal ratio mask (IRM) and the auxiliary complex spectrum. After obtaining the estimated IRM, i.e., \( \tilde{M} \), the enhanced magnitude is obtained by T-F filtering, i.e., \( |\tilde{X}^*| = \tilde{M}|Y| \), and the recovery phase is extracted from the estimated auxiliary complex spectrum. For R-Net, a complex decoder is set to estimate the residual details of the real and imaginary components of the complex spectrum, respectively.

For convolutive reverberation components, current frame of the spectra is related to long-range historical frames, and thus time sequence modeling is particularly important. The temporal convolutional module (TCM) is widely utilized in speech enhancement algorithms to extract the long-term temporal dependencies [17, 26, 18]. However, local details tend to be blurred with the increase of dilation rate [23]. As for speech dereverberation, the local-range detail is also important to provide information about early reverberation. Therefore, we propose a...
squeezed temporal fusion module (S-TFM), which consists of a cascading of early-late interactive units (ELIU). The internal structure of the ELIU is illustrated in Fig. 2. Compared with the basic unit of the previous TCM, ELIU has mainly two differences. First, the feature dimension is squeezed to a lower embedding space, i.e., 64 by 1 × 1-Conv to effectively mitigate the parameter burden, instead of projecting to a higher one, i.e., 512. Second, two parallel branches are set to model early and late reverberation. Considering that late reverberation can be regarded as a nonlinear additive noise while the early reverberation is linearly correlated with the direct speech, gated linear unit (GLU) [27] and linear convolution are adopted, respectively. Besides, the dilation rates of GLU exponentially increase to capture late reverberation under long RT₆₀ while the dilation rates of linear convolution are constant at 1 to focus on details of early reverberation.

### 2.4. Trade-off effect and loss selection

Inspired by [16], we present detailed mathematical derivations to reveal the trade-off phenomenon between magnitude distortion and phase recovery, which is realized by the linear weighting strategy toward RI loss and magnitude regularization term. The magnitude loss and the RI loss can be, respectively, formulated as

\[ L_{Mag}(X, \tilde{X}) = \left\| X - \tilde{X} \right\|_2^2, \]

\[ L_{RI}(X, \tilde{X}) = \left\| X - \tilde{X} \right\|_2^2 + \left\| X - \tilde{X} \right\|_2^2. \]

The typical linear combination of the magnitude and RI constraint can be generalized as

\[ L_{(\alpha)}(X, \tilde{X}) = \alpha L_{Mag}(X, \tilde{X}) + (1 - \alpha) L_{RI}(X, \tilde{X}), \]

where \( \alpha \) is the linear weighted coefficient. Let us take the partial derivative of \( L_{(\alpha)} \) with respect to \( \tilde{X} \), which can be given as

\[ \frac{\partial L_{(\alpha)}(X, \tilde{X})}{\partial \tilde{X}} = \frac{\partial}{\partial \tilde{X}} \left[ \alpha L_{Mag}(X, \tilde{X}) + (1 - \alpha) L_{RI}(X, \tilde{X}) \right] \]

\[ = 2\tilde{X} - 2X \left[ (1 - \alpha) \cos(\Delta \theta) + \alpha \right], \]

where \( \Delta \theta = \theta_X - \theta_{\tilde{X}} \) refers to the phase difference between the estimated \( \theta_{\tilde{X}} \) and target \( \theta_X \). Let \( \frac{\partial L_{(\alpha)}(X, \tilde{X})}{\partial \tilde{X}} = 0 \) and the theoretical optimal magnitude estimation \( \bar{X} \) under \( \Delta \theta \) is

\[ \bar{X}_{opt(\alpha)} = \max \{ X \mid (1 - \alpha) \cos(\Delta \theta) + \alpha \} \geq 0, \]

where \( \max \{ , 0 \} \) is to guarantee the non-negativity of \( X \).

From Eq. (11) one can see that \( \alpha \) weighs the tendency to enhance magnitude with less magnitude distortion or to emphasize phase constraint to recover accuracy phase. Therefore, for the loss function utilized for the magnitude estimation stream in the first stage, i.e., \( L_{Mag} \), \( \alpha = 1 \) is set to solely constrain the magnitude optimization space, which can theoretically obtain the distortionless estimated magnitude while maintaining the phase unaltered. On the contrary, for the loss function utilized for the phase recovery stream, i.e., \( L_{Pha} \), \( \alpha = 0 \) is set to greatly emphasize phase constraints and recover phase information as much as possible. To simultaneously optimize these two individual tasks in the first stage, an equal-weighted linear sum of \( L_{Mag} \) and \( L_{Pha} \) is set. As for the loss function used for the second stage, i.e., \( L_{sz} \), an empirical compromised coefficient, i.e., \( \alpha = 0.5 \), is selected to balance magnitude estimation refinement and phase recovery refinement. To summarize, the loss functions in each stage are written as follows

\[ L_{Mag} = L_{(1)}(X, |\tilde{X}^s|), \]

\[ L_{Pha} = L_{(0)}(X, \tilde{X}^{w2r}), \]

\[ L_{sz} = \frac{1}{2} \{ L_{Mag} + L_{Pha} \}, \]

\[ L_{sz} = L_{(0.5)}(X, \tilde{X}^s). \]

### 3. Experimental settings

#### 3.1. Dataset setup

Experiments are conducted on the English reading speech of the DNS-Challenge dataset [28], which is derived from Librivox, consisting of totally 65,348 clean clips. 45,000, and 3,000 clips are randomly sampled for training and validation, respectively. 150 utterances from the remaining untrained speakers are selected for model evaluation.

In this study, 28 various rooms are firstly simulated to generate a set of RIRs. The length-width-height of the simulated rooms are randomly sampled from 1 m-1 to 10 m-10 m-5 m. In each room, we randomly select the receiver and source positions, and the RT₆₀ is uniformly sampled from 0.3 s to 1.0 s, 100 RIRs are generated with the image method [29]. As a result, there are totally 2,800 RIRs generated for training. Following Eq. (1), we convolve the clean utterance with a randomly selected RIR to generate the reverberant speech, and the corresponding direct-path speech is chosen as the target. In total, 45,000, and 3,000 reverberant-direct pairs are generated for training and validation, respectively.

For model evaluation, another 150 RIRs with untrained room configurations are generated under three RT₆₀ ∈ \{0.3, 0.6, 0.9 s\} and 150 reverberant-direct pairs are generated for each case. We denote this test set as Test A. To test if the trained model can generalize to real RIR conditions, another three test sets are generated by convolving the real measured RIRs taken from the REVERB Challenge [5] with unseen clean speech, namely small room, medium room, and large room test set. 150 reverberant-direct pairs are generated for each case and this test set is denoted as Test B.

#### 3.2. Baseline systems

We compare our proposed framework with another four advanced baselines, namely LSTM [9], CRN [30], GCRN [12], and a two-stage model, i.e., CTS-Net [18]. For LSTM, two LSTM layers with 1024 units are stacked, followed by a dense layer with a ReLU activation function to generate the magnitude. CRN is a typical encoder-decoder topology with two LSTM layers for sequence modeling to estimate the magnitude. GCRN is an improved version of CRN to estimate complex spectrum of target speech, where all the (de)convolution layers are replaced by the ConvGLUs. CTS-Net is a two-stage model, where the first stage generates the magnitude and the second stage refines the complex spectrum. The network for each stage is also an encoder-decoder topology but eighteen light-weight TCMs are utilized for sequence modeling. All the models are trained with the best configurations mentioned in the literature.

#### 3.3. Parameter configurations

The detailed parameter configurations of our system are listed as follows. Both B-Net and R-Net consist of five convolutional blocks in the encoders, and each convolutional block includes a ConvGLU layer, instance normalization (IN), and PReLU [23]. The kernel size and stride are (2, 3) and (1, 2) in the time and frequency axis, respectively. The decoder is the mirror version of the encoder, except all the convolution layers are replaced
Table 1: Average PESQ and ESTOI scores on Test A and Test B among different speech dereverberation models. Best records are highlighted in **BOLD**.

<table>
<thead>
<tr>
<th>Systems</th>
<th>Test A</th>
<th>Test B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PESQ</td>
<td>ESTOI (%)</td>
</tr>
<tr>
<td></td>
<td>0.9 s</td>
<td>0.6 s</td>
</tr>
<tr>
<td>Reverberant</td>
<td>1.79 ± 2.01</td>
<td>2.57 ± 2.12</td>
</tr>
<tr>
<td>LSTM</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CRN</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>GCRN</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CTS-Net</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>B-Net</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Proposed</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 2: Ablation study for first stage module and sequence modeling unit on Test A.

<table>
<thead>
<tr>
<th>Models</th>
<th>Id</th>
<th>First Stage</th>
<th>Second Stage</th>
<th>Sequence Module</th>
<th>PESQ</th>
<th>ESTOI (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reverberant</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.12</td>
<td>51.32</td>
</tr>
<tr>
<td>Proposed</td>
<td>1</td>
<td>Mag.-branch</td>
<td>R-Net</td>
<td>S-TFM</td>
<td>2.68</td>
<td>73.76</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Pha.-branch</td>
<td>R-Net</td>
<td>S-TFM</td>
<td>2.67</td>
<td>74.40</td>
</tr>
<tr>
<td>Proposed</td>
<td>3</td>
<td>B-Net</td>
<td>R-Net</td>
<td>S-TCM</td>
<td>2.76</td>
<td>75.92</td>
</tr>
<tr>
<td>Proposed</td>
<td>4</td>
<td>B-Net</td>
<td>R-Net</td>
<td>S-TCM</td>
<td>2.78</td>
<td>76.19</td>
</tr>
</tbody>
</table>

4. Results and analysis

4.1. Objective comparison results

Two objective metrics are employed to evaluate speech quality and intelligibility, namely perceptual evaluation speech quality (PESQ) [33], and extended short-time objective intelligibility (ESTOI) [34]. The higher the values, the better the performance. The quantitative results on Test A are presented in Table 1 and several observations can be made. Firstly, compared with magnitude-based baselines, i.e., LSTM and CRN, the complex-based baseline, i.e., GCRN, obtains only limited PESQ improvements, indicating that in speech dereverberation tasks, complex spectrum is difficult to be directly recovered by the one-stage topology. Secondly, compared the GCRN with the first stage of our algorithm, i.e., B-Net, around 0.11 and 3.46% improvements are achieved in terms of PESQ and ESTOI, which proves the significance of optimizing magnitude and phase separately. Thirdly, two-stage methods, i.e., CTS-Net, and our proposed method, outperform single-stage baselines and B-Net, indicating the effectiveness of two-stage topology. Fourthly, the proposed algorithm yields state-of-the-art performance in both objective metrics. For example, 0.16 PESQ improvement is achieved from CTS-Net to our approach, demonstrating the superiority of our proposed algorithm.

The similar conclusions can be drawn from Test B in Table 1, which validates the generalization capability of our algorithms in the real reverberated environment. Besides, a subjective AB listening test is conducted and the most preference are given to our proposed algorithm. The result is omitted due to the limited space.

4.2. Ablation analysis

The ablation study is conducted to investigate the effect of the two branches in B-Net and S-TFM, the results are shown in Table 2. A network with only magnitude enhancement branch and a network with only phase recovery branch are involved to be compared with B-Net. For the sequence learning module, S-TFM is compared with the S-TCM [18]. Results indicate that despite the effectiveness of sorely recovering magnitude or phase in the first stage, overall combination of these two branches can further improve speech quality and intelligibility. To give an intuitive interpretation, we plot spectrograms of speeches enhanced by different parts of the proposed approach in Fig. 3. Visualization in Fig. 3 (c) and (g) confirm that the two branches of B-Net can significantly recover magnitude and phase, respectively. And R-Net can further refine the complex spectrum, as shown in Fig. 3 (d) and (h). Besides, S-TFM obtains better performance than S-TCM, verifying the superiority of S-TFM in sequence modeling for speech dereverberation.

5. Conclusions

This paper proposes a novel loss-guided two-stage framework for speech dereverberation. In the first stage, a bifurcated network and well-designed losses are involved in a separate way for efficient phase recovery and magnitude estimation. Entire complex spectrum refinement is conducted by the second stage. The mathematical derivation guides the choice of loss functions theoretically. Experimental results show the superiority of our method over previous advanced methods and ablation studies validate the effectiveness of the proposed framework.
6. References


