

Taylor, Can You Hear Me Now? A Taylor-Unfolding Framework for Monaural Speech Enhancement

Andong Li^{1,2}, Shan You³, Guochen Yu^{1,4}, Chengshi Zheng^{1,2*}, Xiaodong Li^{1,2}

¹Institute of Acoustics, Chinese Academy of Sciences

²University of Chinese Academy of Sciences

³SenseTime Research

⁴Communication University of China

{liandong, cszheng, lxd}@mail.ioa.ac.cn, youshan@sensetime.com, yuguochen@cuc.edu.cn

Abstract

While the deep learning techniques promote the rapid development of the speech enhancement (SE) community, most schemes only pursue the performance in a black-box manner and lack adequate model interpretability. Inspired by Taylor’s approximation theory, we propose an interpretable decoupling-style SE framework, which disentangles the complex spectrum recovery into two separate optimization problems i.e., magnitude and complex residual estimation. Specifically, serving as the 0th-order term in Taylor’s series, a filter network is delicately devised to suppress the noise component only in the magnitude domain and obtain a coarse spectrum. To refine the phase distribution, we estimate the sparse complex residual, which is defined as the difference between target and coarse spectra, and measures the phase gap. In this study, we formulate the residual component as the combination of various high-order Taylor terms and propose a lightweight trainable module to replace the complicated derivative operator between adjacent terms. Finally, following Taylor’s formula, we can reconstruct the target spectrum by the superimposition between 0th-order and high-order terms. Experimental results on two benchmark datasets show that our framework achieves state-of-the-art performance over previous competing baselines in various evaluation metrics. The source code is available at <https://github.com/Andong-Li-speech/TaylorSENet>.

1 Introduction

As a consequence of the COVID-19 pandemic, people and organizations have become increasingly dependent on remote communication techniques to stay connected and conduct business routines. It is thus imperative to demand high-quality speech when background noise and room reverberation exist. As a resolution, monaural speech enhancement (SE) aims to extract the target speech from the noisy mixture when only the single-channel recording is available.

*Corresponding Author

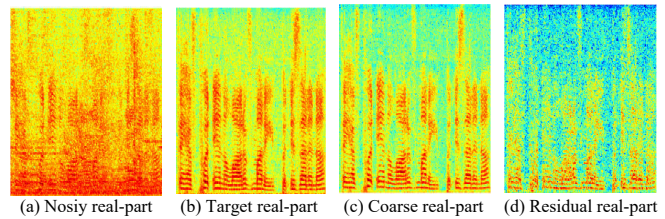


Figure 1: Example of the decoupling target, and only the real part is visualized for illustration convenience.

Recently, the advent of deep neural networks (DNNs) has significantly promoted the performance of SE algorithms, which can be roughly categorized into two streams, namely in the time-domain [Pascual *et al.*, 2017; Luo and Mesgarani, 2019] and time-frequency (T-F)-domain [Yin *et al.*, 2020; Tang *et al.*, 2021]. As speech and noise patterns tend to be more distinguishable after the short-time Fourier transform (STFT), the latter still dominates the mainstream. Previous works simply estimated the magnitude of the spectrum and left the noisy phase unaltered, but they would inevitably incur heavy performance restrictions. To address this problem, it is necessary to consider the joint optimization of magnitude and phase. In [Yin *et al.*, 2020], a dual-branch network was designed to separately model the magnitude and cosine representations of phase. However, due to the nonstructural characteristic of phase, the phase-branch tends to be sensitive to nonlinear operations. Another typical strategy is to couple the magnitude and phase into Cartesian coordinates and construct real and imaginary (RI) pairs. By virtue of complex spectral mapping (CSM) [Tan and Wang, 2020] or complex ratio mask (CRM) [Williamson *et al.*, 2015], both the magnitude and phase can be implicitly recovered. However, such target entanglement will cause the compensation effect [Wang *et al.*, 2021], i.e., magnitude distortion is inevitably sacrificed to compensate for the phase prediction accuracy, especially under low signal-to-noise ratios (SNRs).

More recently, a decoupling mapping procedure [Li *et al.*, 2021] was proposed to decouple the complex spectrum estimation into two separate steps. In the first step, only the magnitude prior is estimated, which is coupled with the noisy phase to obtain a coarse complex spectrum. Afterward, with residual learning, another network is utilized to estimate the

residual component with sparse distribution in the complex domain, which measures the gap between noisy and target phases. Different from previous literature, it endows the separate optimization space toward the magnitude and phase, and therefore, alleviates the compensation effect. Figure 1(c)-(d) visualize the coarse and residual spectra as an example.

In this paper, we rethink the spectrum decoupling and formulate it as an approximation problem *w.r.t.* input neighborhood space. In other words, *if we can access the complex residual and repair the phase in advance, then we can perfectly approximate the clean spectrum from magnitude estimation theoretically*. This process can be expressed as $S = \mathcal{F}(X + \delta)$, where $\{S, X, \delta\}$ denote clean, mixture, and residual components, respectively, and \mathcal{F} is the magnitude estimation function. Based on this new formulation, it is intuitive to leverage Taylor’s approximation *w.r.t.* X to estimate the function representation at $X + \delta$. However, in practical implementation, the residual prior is usually inaccessible. In this regard, we propose a new framework called **TaylorSENet** to explicitly model the Taylor’s approximation by formulating the main term and derivative term as learnable modules. Concretely, the 0th-order module is to consider the magnitude of the spectrum while the high-order modules are concerned with complex residual estimation. In this way, the estimation of the complex spectrum can be obtained by the superimposition of the 0th-order non-derivative and multiple high-order derivative terms. Different from previous SE models in a black-box manner, the proposed framework provides each module with better interpretability. To the best of our knowledge, this is the first time to cast the complex spectrum recovery as Taylor’s approximation problem in the speech front-end task. Our contributions can be summarized as three-fold:

- We rethink the decoupling-style SE algorithm and abstract it as Taylor’s approximation problem.
- We propose an end-to-end framework to simulate the 0th-order and high-order items of Taylor-unfolding.
- We conduct comprehensive experiments and the results show that our system achieves state-of-the-art performance among two benchmarks.

2 Related Work

In this section, we first review the SE task from the T-F domain and time domain. Then we briefly illustrate the multi-stage learning paradigm, which has gradually arisen and gained superiority in the speech front-end fields more recently.

T-F Domain Methods. Before entering the deep learning era, traditional denoising algorithms were applied in the T-F domain as Fourier theory provides a feasible feature representation. Typical methods include spectral subtraction [Boll, 1979], Wiener filtering [Scalart and others, 1996], and statistical-based methods [Ephraim and Malah, 1984]. After the proliferation of DNNs, the denoising task is formulated into a supervised learning problem and can be trained to grasp the latent mapping relations between noisy features

and clean targets. For a long time, only the magnitude is considered as the phase distribution is nonstructured and is thus difficult to predict. More recently, increasing evidence shows that phase also plays a pivotal role in perception quality improvement [Paliwal *et al.*, 2011].

For phase-aware SE approaches, in GCRN [Tan and Wang, 2020], a UNet-style network was utilized to estimate both real and imaginary (RI) parts of the spectrum. As a modification, DCCRN [Hu *et al.*, 2020] devised a complex-valued UNet to adapt the complex correlation between RI. [Yin *et al.*, 2020] proposed a dual-branch network to model the magnitude filter and cosine representations of the phase, respectively. CT-SNet [Li *et al.*, 2021] proposed a two-stage mapping regime, where the magnitude was first estimated as the prior to further facilitate the subsequent phase recovery.

Time Domain Methods. Thanks to the development of DNNs, time-domain-based methods have gained prosperity more recently. A typical option is to directly learn the sample distribution. In DDAEC [Pandey and Wang, 2020], the 1-D waveform was first enframed into 2-D format and then passed through a UNet structure with layer-wise dense-nets. SEGAN [Pascual *et al.*, 2017] adopted the generative adversarial network (GAN) to predict the waveform directly. Another strategy is to use a pair of learnable encoder and decoder to convert the waveform samples into latent space and then devise a separate module to distinguish between different sources. Representative works include Conv-TasNet [Luo and Mesgarani, 2019] and DPRNN [Luo *et al.*, 2020].

Multi-stage Learning. Due to the lack of prior information, the performance of a single-stage SE pipeline is heavily limited in complicated acoustic scenarios. In contrast, in a multi-stage pipeline, the original mapping problem is usually decomposed into several separate subtasks and enables progressive learning. Besides, the previous estimation also serves as the latent prior to guide the subsequent learning process. [Westhausen and Meyer, 2020] combined the complementarity of T-F and latent domain and proposed a stacked dual signal transformation network. [Hao *et al.*, 2021] rethought spectrum recovery from the dual optimization of subband and fullband and proposed a two-stage approach to capture both local and global spectral contexts. Based on the glance and gaze behavior of humans in visual perception, [Li *et al.*, 2022] proposed to stack multiples glance-gaze modules to reconstruct the spectrum collaboratively.

3 Methodology

3.1 From Decoupling to Taylor’s Series Expansion

Given STFT, the mixture signal in the T-F domain can be formulated as:

$$X_{k,l} = S_{k,l} + N_{k,l}, \quad (1)$$

where $\{X_{k,l}, S_{k,l}, N_{k,l}\} \in \mathbb{C}$ denote complex-valued noisy, clean, and noise signals in the frequency bin of $k \in \{1, \dots, K\}$ and time index of $l \in \{1, \dots, L\}$. For brevity, we omit the subscripts $\{k, l\}$ if no confusion arises. The aim of SE is to design an operator to extract the target speech from the noisy mixture, *i.e.*,

$$S = \mathcal{F}_1(X), \quad (2)$$

where \mathcal{F}_1 denotes the estimation function. Although various networks can be employed to accomplish this process, these methods usually encapsulate the whole recovery process as a black-box and thus have weak interpretability in the intermediate stages [Tan and Wang, 2020; Hu *et al.*, 2020]. Recently, a decoupling-style forward pipeline was proposed in [Li *et al.*, 2021] and given as

$$|S| = \mathcal{F}_{mag}(|X|), \quad (3)$$

$$S = |S| e^{j\theta_X} + \mathcal{F}_{com}(S, X), \quad (4)$$

where \mathcal{F}_{mag} and \mathcal{F}_{com} are the mapping functions for magnitude and residual estimation, respectively, and θ_X denotes the noisy phase. The above procedure decomposes the entanglement of magnitude and phase by step-wise optimization, and can be summarized into two core operations:

- **op1**: suppress the noise in the magnitude domain while ignoring the phase term to obtain a coarse estimation.
- **op2**: estimate the complex residual while fixing the magnitude term to refine the target spectrum.

In traditional SE algorithms, **op1** can be fulfilled by either noise subtraction [Boll, 1979] or noise filtering [Ephraim and Malah, 1984]. For both two techniques, the aforementioned procedure can be respectively expressed as

$$S = (|X| - |N|) e^{j\theta_X} + \delta_1 = X - \tilde{N}_1 + \delta_1, \quad (5)$$

$$S = M |X| e^{j\theta_X} + \delta_2 = X - \tilde{N}_2 + \delta_2, \quad (6)$$

where $\{\tilde{N}_1 = |N| e^{j\theta_X}, \tilde{N}_2 = (1 - M) |X| e^{j\theta_X}\}$ respectively denote the estimated noise components in traditional noise subtraction and noise filtering algorithms, and M denotes the spectral filter gain. $\{\delta_1, \delta_2\}$ denote the corresponding phase residual term.

Comparing Eqn.(5) and Eqn.(6), we can find that they have a similar format, which is intuitive since **op1** essentially encourages an accurate noise power spectral density (NPSD) estimation and then subtracts it from the mixture. Furthermore, denoting $\mathcal{F}(X) = X - \tilde{N}$, and $X := X + \delta$, Eqns.(5)-(6) can thus be abstracted into a more general case:

$$S = \mathcal{F}(X + \delta). \quad (7)$$

Eqn.(7) implies that if we can access the residual term δ and add it into input X in advance, then we are able to perfectly recover the spectrum by magnitude estimation theoretically. However, in practical scenarios, we usually can not access the residual prior δ . Therefore, to resolve the above generalized function, we expand Eqn.(7) with an infinite Taylor's series expansion at X , given as

$$S = \mathcal{F}(X) + \frac{1}{1!} \frac{\partial \mathcal{F}(X)}{\partial X} \delta + \dots + \frac{1}{q!} \frac{\partial^q \mathcal{F}(X)}{\partial^q X} \delta^q + \dots \quad (8)$$

which can be simplified as

$$S = \underbrace{\mathcal{F}(X)}_{\text{0th-order term}} + \underbrace{\sum_{q=1}^{+\infty} \frac{1}{q!} \frac{\partial^q \mathcal{F}(X)}{\partial^q X} \delta^q}_{\text{high-order terms}}. \quad (9)$$

As such, we provide a novel perspective of complex spectrum recovery from Taylor's approximation theory. In detail, the 0th-order term is grounded with regard to the estimation of spectral magnitude, and the high-order terms attempt to approximate the distribution of the residual component.

3.2 Taylor-Unfolding Framework

To adapt the formulation of Taylor's series expansion to the network, it is necessary to derive the correlation between adjacent derivative terms. For practical implementation, we truncate the number of orders in the derivative part into Q and neglect higher-order parts. In light of [Fu *et al.*, 2021], we first define the q th order derivative term as $\mathcal{T}(q, X, \delta)$

$$\mathcal{T}(q, X, \delta) = \frac{\partial^q \mathcal{F}(X)}{\partial^q X} \delta^q, \quad (10)$$

where the factorial term is dropped for derivation convenience. To investigate the correlation between q th order $\mathcal{T}(q, X, \delta)$ and $(q + 1)$ th order $\mathcal{T}(q + 1, X, \delta)$, we differentiate $\mathcal{T}(q, X, \delta)$ w.r.t. X

$$\frac{\partial \mathcal{T}(q, X, \delta)}{\partial X} = \frac{\partial}{\partial X} \left(\frac{\partial^q \mathcal{F}(X)}{\partial^q X} \right) \delta^q + \frac{\partial^q \mathcal{F}(X)}{\partial^q X} \frac{\partial}{\partial X} \delta^q. \quad (11)$$

Considering

$$\frac{\partial}{\partial X} \left(\frac{\partial^q \mathcal{F}(X)}{\partial^q X} \right) = \frac{\partial^{q+1} \mathcal{F}(X)}{\partial^{q+1} X}, \quad (12)$$

$$\frac{\partial}{\partial X} \delta^q = \frac{\partial \delta^q}{\partial \delta} \frac{\partial \delta}{\partial X} = -q \delta^{q-1}. \quad (13)$$

Substituting Eqns.(12)-(13) into Eqn.(11) and multiplying δ on both sides, we can derive the following recursive formula between $\mathcal{T}(q + 1, X, \delta)$ and $\mathcal{T}(q, X, \delta)$:

$$\mathcal{T}(q + 1, X, \delta) = q \mathcal{T}(q, X, \delta) + \delta \frac{\partial \mathcal{T}(q, X, \delta)}{\partial X}. \quad (14)$$

Generally speaking, it is quite difficult to access the derivative operator $\delta \frac{\partial \mathcal{T}(q, X, \delta)}{\partial X}$. To this end, we design a trainable network module, notated as $\mathcal{P}(q, X, \delta)$, to replace the complicated operator. Note that as the network weights are purely learned from the training data, it does not necessarily follow a strict mathematical definition of derivation, but we empirically find that it indeed involves the residual estimation.

3.3 Framework Structure

Forward Stream

We instantiate the proposed Taylor-unfolding framework, as shown in Figure 2(a). It mainly comprises two parts, namely the 0th-order module and multiple high-order modules. According to our previous formulation, the 0th-order module targets at magnitude estimation. To this end, we first convert the input RI into magnitude, *i.e.*, $|X| = \sqrt{X_r^2 + X_i^2}$, and then we sent it to the 0th-order module to obtain the output gain M with range $(0, 1)$ for noise filtering in the magnitude domain, as shown in Eqn.(15). To model high-order terms, we employ the high-order encoder to directly extract the patterns from the RI input, and the output feature map R is then concatenated with the output from the last high-order

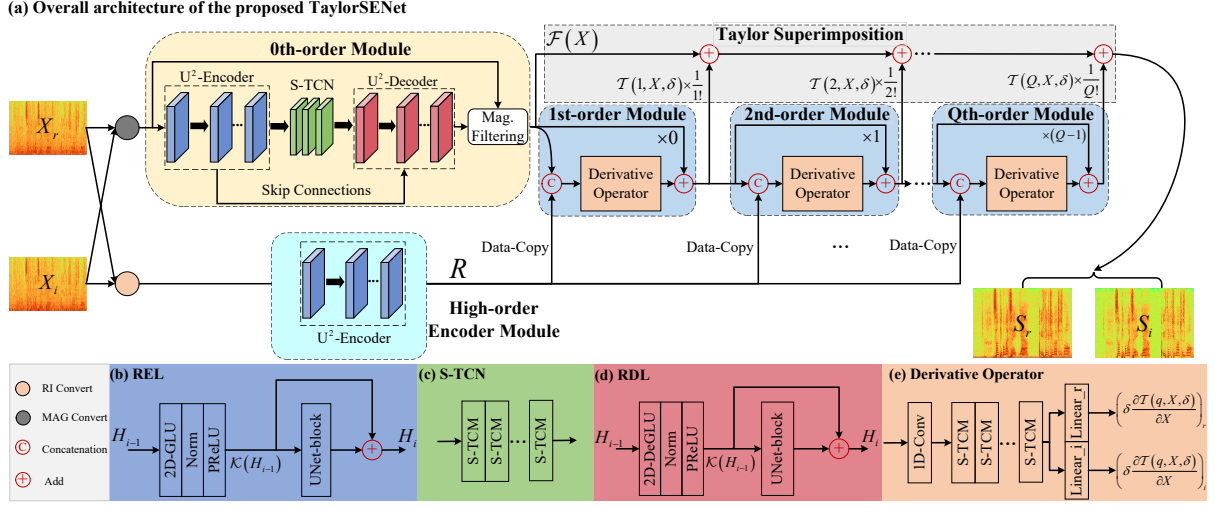


Figure 2: An overview of the proposed TaylorSENet. (a) Network diagram of TaylorSENet, which involves the superimposition of multiple-order terms. The 0th-order module adopts the magnitude filtering. The calculation of high-order terms follows the recursive formula in Eqn.(14). (b) Detail of the recalibration encoding layer (REL) in the U^2 encoder. (c) Detail of the squeezed-TCN (S-TCN). (d) Detail of the recalibration decoding layer (RDL) in the U^2 decoder. (e) Detail of the derivative operator.

module as the input of the next module for high-order term update, as shown in Eqns.(16)-(17). This operation is implemented recursively. Note that after magnitude filtering, we couple the filtered spectral magnitude with the noisy phase θ_X to yield the coarse complex spectrum. After all the terms are obtained, following Taylor’s formula, we superimpose all of them to reconstruct the target spectrum. In a nutshell, the overall forward stream is formulated as:

$$\mathcal{F}(X) = M|X|e^{j\theta_X}, \quad (15)$$

$$\mathcal{P}(q, X, \delta) = \mathcal{G}(\text{Concat}(\mathcal{T}(q, X, \delta), R)), \quad (16)$$

$$\mathcal{T}(q+1, X, \delta) = q \odot \mathcal{T}(q, X, \delta) + \mathcal{P}(q, X, \delta), \quad (17)$$

$$S = \mathcal{F}(X) + \sum_{q=1}^Q \frac{1}{q!} \mathcal{T}(q, X, \delta), \quad (18)$$

where $q \in \{1, \dots, Q\}$ is the order index, and \mathcal{G} denotes the mapping function of derivative operator in the high-order module. \odot denotes the Hadamard product.

0th-Order Module

In the 0th-order module, we adopt a classical UNet-style encoder-decoder structure, which has been widely utilized in the SE task [Tan and Wang, 2020; Li *et al.*, 2021]. The encoder is to gradually decrease the feature size with consecutive downsampling operations while extracting the spectral features. In contrast, the decoder has a mirror structure and attempts to recover the original spectral size with deconvolution layers. Nonetheless, high-level semantic information is usually embedded in the various-length frame correlations and naive convolution operations often can not capture such complicated multiscale information. Inspired by the success of U^2 -Net in the salient object detection field [Qin *et al.*, 2020], we adapt the U^2 -Encoder and U^2 -decoder with multiple recalibration encoding/decoding layers (REL/RDLs) herein, as shown in Figure 2(b)(d). Taking REL as an exam-

ple, each 2D-gated linear unit (GLU) [Dauphin *et al.*, 2017] is followed by instance normalization (IN), and PReLU. Then a UNet-block is inserted with the residual connection, which takes the UNet-style structure except that the layer depth dynamically varies with the current input size. The process can be expressed as:

$$\mathcal{K}(H_{i-1}) = \text{PReLU}(\text{Norm}(2\text{D-GLU}(H_{i-1}))), \quad (19)$$

$$H_i = \mathcal{K}(H_{i-1}) + \text{UNet-block}(\mathcal{K}(H_{i-1})), \quad (20)$$

where H_i denotes the input feature map of the i -th REL. The rationale is two-fold. First, by further feature downsampling-upsampling operations, different scale information can be effectively grasped. Besides, as the feature map close to the input tends to be rather noisy, we can recalibrate the feature map and preserve the target information.

To establish long-term relations between adjacent frames, we stack multiple temporal convolution networks (TCNs) [Luo and Mesgarani, 2019], which comprise multiple temporal convolution modules (TCMs) with increasing dilations in the time axis. Besides, to decrease the computational footprint, we adopt a squeezed version of TCN (dubbed S-TCN) [Li *et al.*, 2021], as shown in Figure 2(c), where the squeezed TCM (S-TCM) is leveraged for more compact channels. We also investigate other advanced structures for 0th-order module such as transformers [Vaswani *et al.*, 2017] and conformers [Gulati *et al.*, 2020] in experimental ablation studies (see Section 4.4).

High-Order Module

For effective network training, we model the complicated derivative operator with a trainable network module, whose internal structure is presented in Figure 2(e). As Eqn.(14) indicates, the operator involves both the last-order term and input X , so we utilize both the encoded feature R from noisy input X and the estimation from the last-order term

| Entry | Param. (M) | Is-shared | Q | Zero-type | PESQ \uparrow | ESTOI(%) \uparrow | SISNR(dB) \uparrow |
|-------|-------------|--------------|-----|----------------|-----------------|---------------------|----------------------|
| 1a | 2.19 | \times | 0 | U ² | 2.63 | 69.66 | 8.48 |
| 1b | 3.76 | \times | 1 | U ² | 2.76 | 74.38 | 10.25 |
| 1c | 4.58 | \times | 2 | U ² | 2.80 | 75.35 | 10.54 |
| 1d | 5.40 | \times | 3 | U ² | 2.81 | 76.05 | 10.79 |
| 1e | 6.22 | \times | 4 | U ² | 2.80 | 75.66 | 10.66 |
| 1f | 7.04 | \times | 5 | U ² | 2.80 | 76.09 | 10.85 |
| 2c | 3.76 | \checkmark | 2 | U ² | 2.77 | 74.90 | 10.34 |
| 2d | 3.76 | \checkmark | 3 | U ² | 2.77 | 74.66 | 10.41 |
| 2e | 3.76 | \checkmark | 4 | U ² | 2.80 | 75.12 | 10.55 |
| 2f | 3.76 | \checkmark | 5 | U ² | 2.77 | 74.74 | 10.29 |
| 3a | 3.48 | \times | 3 | U | 2.68 | 72.93 | 10.09 |
| 3b | 5.70 | \times | 3 | Transformer | 2.62 | 71.92 | 9.85 |
| 3c | 12.85 | \times | 3 | Conformer | 2.67 | 72.93 | 10.15 |

Table 1: Ablation study on WSJ0-SI84 dataset.

$\mathcal{T}(q-1, X, \delta)$ as the input and send to a 1D-Conv. Several S-TCMs are employed as the modeling unit, and we can generate $\delta \frac{\partial \mathcal{T}(q, X, \delta)}{\partial X}$ with linear transformation. Note that our framework also applies to other more advanced network structures and are expected to achieve even better performance, which we leave as future work. Besides, as the derivative operator should be parameter-invariant theoretically, we also investigate the case when the parameters are shared among different derivative modules (see Section 4.4).

4 Experiments

4.1 Datasets

WSJ0-SI84. It consists of 7138 utterances by 83 speakers (42 males and 41 females) [Paul and Baker, 1992]. We randomly select 5428 and 957 clips for training and validation, and another 150 clips by untrained speakers are used for testing. To generate noisy-clean pairs, we sample around 20,000 noises from the DNS-Challenge noise set [Reddy *et al.*, 2020] and the training SNRs are sampled from [-5dB, 0dB]. As a result, approximately 300 hours pairs are created for training. For model evaluation, three untrained challenging noises are selected, namely babble, factory1, and cafeteria. Three testing SNRs are set, *i.e.*, $\{-3\text{dB}, 0\text{dB}, 3\text{dB}\}$, and 150 pairs are created for each case.

DNS-Challenge. The Interspeech 2020 DNS-Challenge corpus covers over 500 hours of clean clips by 2150 speakers and over 180 hours of noise clips [Reddy *et al.*, 2020]. For model evaluation, it provides a non-blind validation set with two categories, namely with and without reverberation, and each includes 150 noisy-clean pairs. Following the scripts provided by the organizer, we generate around 3000 hours of noisy-clean pairs for training and the SNRs randomly range from -5dB to 15dB.

4.2 Configuration Details

Network configuration. In the U²-encoder and U²-decoder, the kernel size and stride of 2D-GLUs are namely set as (1, 3) and (1, 2) in the time and frequency axes, and we set the kernel size in the UNet-block as (2, 3). The number

| Methods | Do. | Param. (M) | MACs (G/s) | RTF | NB-PESQ \uparrow | ESTOI \uparrow (%) | SISNR \uparrow (dB) |
|--------------------------|-----|-------------|-------------|-------------|--------------------|----------------------|-----------------------|
| Noisy | - | - | - | - | 1.82 | 41.97 | 0.00 |
| GCRN [2020] | T-F | 9.77 | 2.42 | 0.21 | 2.48 | 70.68 | 9.21 |
| DCCRN [2020] | T-F | 3.67 | 11.13 | 0.44 | 2.54 | 70.58 | 9.47 |
| PHASEN [2020] | T-F | 8.76 | 6.12 | 0.37 | 2.73 | 71.77 | 9.38 |
| FullSubNet [2021] | T-F | 5.64 | 31.35 | 1.35 | 2.55 | 65.89 | 9.16 |
| CTSNNet [2021] | T-F | 4.35 | 5.57 | 0.41 | 2.86 | 76.15 | 10.92 |
| GaGNet [2022] | T-F | 5.94 | 1.63 | 0.10 | 2.86 | 76.87 | 10.93 |
| Conv-TasNet [2019] | T | 4.35 | 5.57 | 0.73 | 2.48 | 72.01 | 9.99 |
| DPRNN [2020] | T | 4.22 | 9.20 | 0.77 | 2.54 | 73.18 | 10.21 |
| DDAEC [2020] | T | 4.82 | 36.56 | 2.40 | 2.76 | 74.84 | 10.85 |
| DEMUCAS [2020] | T | 18.87 | 4.35 | 0.28 | 2.67 | 76.23 | 11.08 |
| TaylorSENet(Ours) | T-F | 3.76 | 6.14 | 0.50 | 2.89 | 76.95 | 10.82 |
| TaylorSENet(Ours) | T-F | 5.40 | 6.14 | 0.51 | 2.91 | 77.76 | 11.16 |

Table 2: Quantitative comparisons with other SOTA systems on WSJ0-SI84 dataset. Scores are averaged upon different testing cases. ‘‘Do.’’ denotes the transform domain of the method.

of 2D-conv channels remains 64 by default. Denote the number of (de)encoder layers in the i -th UNet-block as U_i , and then $U = \{4, 3, 2, 1, 0\}$ where 0 means no UNet-block is employed. For S-TCN and derivative operators, similar to [Li *et al.*, 2022], two groups of S-TCMs are utilized, each of which includes four S-TCMs with kernel size and dilation rates of 5 and $\{1, 2, 5, 9\}$, respectively. Causal convolution operations are adopted by zero-padding along the past frames.

Training configuration. We sample all the utterances at 16 kHz. The window size is set as 20 ms, with 50% overlap between adjacent frames. 320-point FFT is utilized, leading to 161-D in the feature axis. The model is trained on Pytorch platform with a NVIDIA V100 GPU. We use the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) with a batch size of 8 to train the proposed model, and the learning rate is initialized as 5e-4. For WSJ0-SI84, we train the model by 60 epochs, while 30 epochs for DNS-Challenge. The RI loss together with magnitude constraint is utilized as the training loss and the power-spectrum compression strategy is utilized with the compression coefficient empirically set as 0.5 [Li *et al.*, 2022]. The learning rate will be halved if the validation loss does not decrease for two consecutive epochs.

4.3 Evaluation Metrics

Multiple objective metrics are adopted, including narrow-band (NB) [Rix *et al.*, 2001] and wide-band (WB) perceptual evaluation speech quality (PESQ) [Rec, 2005] for speech quality, short-time objective intelligibility (STOI) [Taal *et al.*, 2011] and its extended version ESTOI [Jensen and Taal, 2016] for intelligibility, and SISNR [Roux *et al.*, 2019] for speech distortion. For all the metrics, high values indicate better performance.

4.4 Ablation Study

We randomly sample around 100-hour pairs from WSJ0-SI84 corpus to conduct the ablation study, which spans the following three aspects: (1) Is-shared: whether the parameters are shared among multiple high-order modules; (2) Q : the number of derivative orders; and (3) Zero-type: network type used in the 0th-order module. We fix the random seed, and

| Methods | Do. | w/ Reverberation | | | | w/o Reverberation | | | |
|--------------------------|-----|------------------|-------------|--------------|--------------|-------------------|-------------|--------------|--------------|
| | | WB-PESQ | NB-PESQ | STOI(%) | SISNR (dB) | WB-PESQ | NB-PESQ | STOI(%) | SISNR(dB) |
| Noisy | - | 1.82 | 2.75 | 86.62 | 9.03 | 1.58 | 2.45 | 91.52 | 9.07 |
| NSNet [2020] | T-F | 2.37 | 3.08 | 90.43 | 14.72 | 2.15 | 2.87 | 94.47 | 15.61 |
| DTLN [2020] | T-F | - | 2.70 | 84.68 | 10.53 | - | 3.04 | 94.76 | 16.34 |
| DCCRN [2020] | T-F | - | 3.32 | - | - | - | 3.27 | - | - |
| FullSubNet [2021] | T-F | 2.97 | 3.47 | 92.62 | 15.75 | 2.78 | 3.31 | 96.11 | 17.29 |
| TRU-Net [2021] | T-F | 2.74 | 3.35 | 91.29 | 14.87 | 2.86 | 3.36 | 96.32 | 17.55 |
| CTS-Net [2021] | T-F | 3.02 | 3.47 | 92.70 | 15.58 | 2.94 | 3.42 | 96.66 | 17.99 |
| GaGNet [2022] | T-F | 3.18 | 3.57 | 93.22 | 16.57 | 3.17 | 3.56 | 97.13 | 18.91 |
| TaylorSENet(Ours) | T-F | 3.33 | 3.65 | 93.99 | 17.10 | 3.22 | 3.59 | 97.36 | 19.15 |

Table 3: Quantitative comparisons with other state-of-the-art systems on the DNS Challenge dataset. “-” denotes no published result.

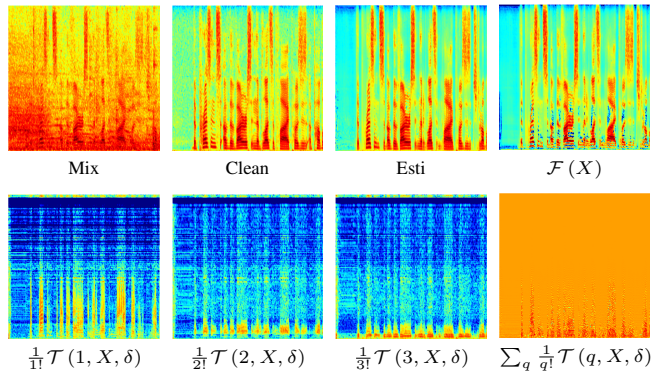


Figure 3: Estimation visualization of intermediate modules.

PESQ, ESTOI, and SISNR are utilized as the evaluation metric, whose results are shown in Table 1.

Effect of the parameter-shared scheme for high-order modeling. In the aforementioned problem formulation, we adopt a trainable network to model the complicated derivative operator for high-order modeling. It is thus necessary to explore whether the parameters can be shared among modules. By comparing the entries from entries 1c-1f to 2c-2f in Table 1, the non-shared case yields relatively better performance over the shared case, especially in terms of ESTOI and SISNR. We can explain the phenomenon from two aspects. First, the former scheme provides more parameter freedom in the optimization space by independent gradient update, while the latter has to balance all the terms with only one set of parameters. Besides, PESQ mainly focuses on the magnitude envelope similarity while ESTOI and SISNR have a close relation to the phase accuracy [Wang *et al.*, 2021], and high-order terms are mainly responsible for phase modification. As a consequence, the nonshared scheme only yields a slightly better PESQ score than the shared case and relatively more notable improvements in terms of ESTOI and SISNR.

Effect of the order number. As shown in entries 1a-1f of Table 2. When $Q = 0$, *i.e.*, no high-order module is employed, it is not surprising to observe the worst performance as only the magnitude is considered and the phase term is neglected. When Q increases from 1 to 3, consistent improvements are achieved among the three metrics, which shows

the effectiveness of Taylor series modeling. However, when Q further increases, the performance inclines to saturate and even slightly drop, *e.g.*, $Q = 4$. A similar trend is also observed in entries 2c-2f. This might result from that high-order modules are responsible for complex residual modeling, which has a rather sparse spectral distribution. Therefore, 3-order can be sufficient to approximate the real distribution.

To further validate the mechanism of Taylor-unfolding framework, taking $Q = 3$ as an example, we visualize the output of each order in Figure 3. Noisy, clean, and final estimated spectra are also presented as reference. As we can see, the output from the 0th-order module is quite similar to the clean version, and most noise components are suppressed, which infers that 0th-order indeed serves as the filter to eliminate the noise and captures the overall speech structure in the magnitude domain. Besides, the outputs in the high-order modules seem rather sparse, but we can notice the contour in the harmonic region, indicating that the high-order module effectively captures the sparse residual structure to refine the phase. Remark that as we only supervise the final estimation, despite not strictly following the mathematical definition of Taylor’s approximation, the network still learns to allocate the role of 0th-order and high-order terms as expected.

Effect of network types in 0th-order module. To validate the superiority of the network used in the 0th-order module, we investigate other networks, as shown in entries 3a-3c of Table 1. “U” denotes that no UNet-block is utilized, and “Transformer” and “Conformer” respectively denote the network is replaced by six transformer and conformer encoding layers, respectively. As we can see, a notable performance drop is observed from entry 1d to 3a, suggesting the effectiveness of UNet-block in feature representation. It is interesting to notice that despite transformer and conformer have shown promising performance in ASR and NLP-related tasks more recently, they are inferior to the proposed U^2 version herein. This might because they mainly consider the global sequential correlations and ignore the local spectral patterns, which hampers the overall filter estimation.

4.5 Comparison with the State-of-the-Art Methods

WSJ0-SI84. Configurations of entries 1d and 2d in Table 1 are selected for comparisons with other ten top-performed baselines, where the superscript “†” denotes that the parameters are shared among high-order modules. Note that ex-

cept PHASEN, all baselines are based on causal implementation. The quantitative results are shown in Table 2, where NB-PESQ, ESTOI, and SISNR are utilized as the evaluation metrics. We also present the number of parameters, multiply-accumulate operations (MACs) per second, and real-time factor (RTF) to evaluate the model computational complexity. As we can see, our method achieves the highest metric scores among all the systems with reasonable trainable parameters and computational complexity. Even with the causal convolutions, our method still dramatically surpasses PHASEN, a noncausal system, which reveals the superiority of our system based on Taylor-unfolding theory. It is interesting to notice that, when parameters are shared, our method will slightly degrade in metric scores, nonetheless, it is also comparable over the state-of-the-art systems but with relatively less trainable parameters.

DNS-Challenge. To verify the superiority of the proposed SE system in more complicated acoustic scenarios, we report the results on Interspeech2020 DNS-Challenge corpus, as shown in Table 3. WB-PESQ, NB-PESQ, STOI, and SISNR are used for evaluation. One can see that our method achieves the highest metric performance in both reverberation and anechoic acoustic scenarios. It substantially demonstrates the denoising potential of our method in both reverberant and anechoic environments.

5 Conclusion

We propose a decoupling-style framework based on Taylor’s approximation theory for speech enhancement. Specifically, the original complex spectrum reconstruction is decoupled into two parts, namely magnitude estimation and complex residual estimation. For the former, a magnitude filter is devised to suppress the noise components in the magnitude domain. For the latter, multiple trainable modules are unfolded to simulate the complicated derivation operator and estimate the corresponding high-order terms. Afterward, we can recover the target spectrum by superimposition following Taylor’s formula. Experiments show that our method achieves state-of-the-art performance over previous top-performed baselines and provides better internal interpretability at the same time.

References

- [Boll, 1979] Steven Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Process.*, 27(2):113–120, 1979.
- [Choi *et al.*, 2021] Hyeong-Seok Choi, Sungjin Park, Jie Hwan Lee, Hoon Heo, Dongsuk Jeon, and Kyogu Lee. Real-Time Denoising and Dereverberation with Tiny Recurrent U-Net. In *ICASSP*, pages 5789–5793, 2021.
- [Dauphin *et al.*, 2017] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *ICML*, pages 933–941, 2017.
- [Defossez *et al.*, 2020] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. In *Interspeech*, pages 3291–3295, 2020.
- [Ephraim and Malah, 1984] Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Process.*, 32(6):1109–1121, 1984.
- [Fu *et al.*, 2021] Xueyang Fu, Zeyu Xiao, Gang Yang, Aiping Liu, Zhiwei Xiong, et al. Unfolding Taylor’s Approximations for Image Restoration. In *NeurIPS*, 2021.
- [Gulati *et al.*, 2020] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech*, pages 5036–5040, 2020.
- [Hao *et al.*, 2021] Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li. Fullsubnet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement. In *ICASSP*, pages 6633–6637, 2021.
- [Hu *et al.*, 2020] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. In *Interspeech*, pages 2472–2476, 2020.
- [Jensen and Taal, 2016] Jesper Jensen and Cees H Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 24(11):2009–2022, 2016.
- [Li *et al.*, 2021] Andong Li, Wenzhe Liu, Chengshi Zheng, Cunhang Fan, and Xiaodong Li. Two Heads are Better Than One: A Two-Stage Complex Spectral Mapping Approach for Monaural Speech Enhancement. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 29:1829–1843, 2021.
- [Li *et al.*, 2022] Andong Li, Chengshi Zheng, Lu Zhang, and Xiaodong Li. Glance and gaze: A collaborative learning framework for single-channel speech enhancement. *Appl. Acoust.*, 187:108499, 2022.
- [Luo and Mesgarani, 2019] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 27(8):1256–1266, 2019.
- [Luo *et al.*, 2020] Yi Luo, Zhuo Chen, and Takuya Yoshioka. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In *ICASSP*, pages 46–50, 2020.
- [Paliwal *et al.*, 2011] Kuldip Paliwal, Kamil Wójcicki, and Benjamin Shannon. The importance of phase in speech enhancement. *Speech. Commun.*, 53(4):465–494, 2011.
- [Pandey and Wang, 2020] Ashutosh Pandey and DeLiang Wang. Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain. In *ICASSP*, pages 6629–6633, 2020.
- [Pascual *et al.*, 2017] Santiago Pascual, Antonio Bonafonte, and Joan Serra. SEGAN: Speech enhancement generative

- adversarial network. In *Interspeech*, pages 3642–3646, 2017.
- [Paul and Baker, 1992] Douglas B Paul and Janet Baker. The design for the wall street journal-based csr corpus. In *Workshop on Speech and Natural Lang.*, pages 357–362, 1992.
- [Qin *et al.*, 2020] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.*, 106:107404, 2020.
- [Rec, 2005] ITUT Rec. P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs. *International Telecommunication Union, CH–Geneva*, 2005.
- [Reddy *et al.*, 2020] Chandan KA Reddy, Vishak Gopal, Ross Culter, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matuskevych, et al. The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. In *Interspeech*, pages 2492–2496, 2020.
- [Rix *et al.*, 2001] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *ICASSP*, pages 749–752, 2001.
- [Roux *et al.*, 2019] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. SDR-half-baked or well done? In *ICASSP*, pages 626–630, 2019.
- [Scalart and others, 1996] Pascal Scalart et al. Speech enhancement based on a priori signal to noise estimation. In *ICASSP*, pages 629–632, 1996.
- [Taal *et al.*, 2011] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, 19(7):2125–2136, 2011.
- [Tan and Wang, 2020] Ke Tan and DeLiang Wang. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, 28:380–390, 2020.
- [Tang *et al.*, 2021] Chuanxin Tang, Chong Luo, Zhiyuan Zhao, Wenxuan Xie, and Wenjun Zeng. Joint time-frequency and time domain learning for speech enhancement. In *IJCAI*, pages 3816–3822, 2021.
- [Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.
- [Wang *et al.*, 2021] Zhong-Qiu Wang, Gordon Wichern, and Jonathan Le Roux. On The Compensation Between Magnitude and Phase in Speech Separation. *IEEE Signal Process. Lett.*, 28:2018–2022, 2021.
- [Westhausen and Meyer, 2020] Nils L Westhausen and Bernd T Meyer. Dual-signal transformation lstm network for real-time noise suppression. In *Interspeech*, pages 2477–2481, 2020.
- [Williamson *et al.*, 2015] Donald S Williamson, Yuxuan Wang, and DeLiang Wang. Complex ratio masking for monaural speech separation. *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, 24(3):483–492, 2015.
- [Yin *et al.*, 2020] Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. PHASEN: A phase-and-harmonics-aware speech enhancement network. In *AAAI*, pages 9458–9465, 2020.