Neural Networks 168 (2023) 508–517

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

CompNet: Complementary network for single-channel speech enhancement

Cunhang Fan [a], Hongmei Zhang [a], Andong Li [b], Wang Xiang [a], Chengshi Zheng [b], Zhao Lv [a,*], Xiaopei Wu [a,*]

[a] *Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China*

[b] *Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, 100190, Beijing, China*

ARTICLE INFO

Keywords:
Speech enhancement
Complementary
Filtering and refining
Time–frequency domain
Time-domain

ABSTRACT

Recent multi-domain processing methods have demonstrated promising performance for monaural speech enhancement tasks. However, few of them explain why they behave better over single-domain approaches. As an attempt to fill this gap, this paper presents a complementary single-channel speech enhancement network (CompNet) that demonstrates promising denoising capabilities and provides a unique perspective to understand the improvements introduced by multi-domain processing. Specifically, the noisy speech is initially enhanced through a time-domain network. However, despite the waveform can be feasibly recovered, the distribution of the time–frequency bins may still be partly different from the target spectrum when we reconsider the problem in the frequency domain. To solve this problem, we design a dedicated dual-path network as a post-processing module to independently filter the magnitude and refine the phase. This further drives the estimated spectrum to closely approximate the target spectrum in the time–frequency domain. We conduct extensive experiments with the WSJ0-SI84 and VoiceBank + Demand datasets. Objective test results show that the performance of the proposed system is highly competitive with existing systems.

1. Introduction

Speech enhancement aims at removing background noise from the noisy mixture to improve the quality and intelligibility of speech. In real life, speech is often polluted by various noises that can seriously degrade the performance of automatic speech recognition (ASR) (Chan, Jaitly, Le, & Vinyals, 2016; Deng et al., 2022; Zhang, Liu, Huang, & Zhao, 2023), speech coding (Kleijn et al., 2021, 2021; Zhang, Zhang, Zhao, & Du, 2023), and hearing aids (Chen, Feng, Zhang, Lu, & Liu, 2022; Koike-Akino et al., 2023; Van Tasell, 1993). Therefore, speech enhancement is crucial for these back-end tasks. Traditional speech enhancement methods, such as Wiener filtering (Guo & Li, 2023), spectral subtraction (Boll, 1979; Liu, Chen, Yang, & Xie, 2022), and computational auditory scene analysis (CASA) (Rouat, 2008), have been proposed in the past few decades. However, these algorithms suffer from significant performance degradation when the noise is highly nonstationary or the signal-to-noise ratio (SNR) becomes relatively low.

We have witnessed the significant performance improvement of the speech enhancement thanks to the rapid development of deep neural networks (DNNs). Based on DNNs, speech enhancement methods can be broadly categorized into three classes: magnitude, time, and complex domains. Among them, both the magnitude and complex domains belong to the time–frequency (T–F) domain category. For the magnitude domain, the time-domain signal is converted into its T–F representation, and the magnitude-related component is then extracted as both the input and target. Depending on the output of magnitude domain networks, these magnitude domain methods can be divided into mapping-based and masking-based. The mapping-based method uses spectral magnitude or its compressed version as output to estimate clean speech (Han, Wang, & Wang, 2014), while masking-based approaches use masks such as ideal binary mask (IBM) (Kim & Lee, 2023), ideal ratio mask (IRM) (Chen, Chen, Zhang, Hu, & Zhao, 2022; Hu, Fan, & Lu, 2022; Li, Li, Li, & Li, 2021; Wang, Narayanan, & Wang, 2014), and ideal magnitude mask (IAM) (Wang & Chen, 2018a; Xie, Xie, Chen, Zhang, & Li, 2022) as outputs. Paliwal, Wójcicki, and Shannon (2011) have demonstrated the significance of accurate phase estimation in perceptual speech quality improvements. However, in traditional speech enhancement methods, phase recovery is usually not considered as the wrapping effect makes the phase distribution unstructured and difficult to estimate, if not impossible, which inevitably limits the performance upper bound of existing speech enhancement methods. To this end, time domain and complex-domain methods are developed to jointly recover magnitude and phase.

Time-domain methods aim to explicitly or latently model the distribution of waveform samples via feature encoding and decoding

* Corresponding authors.
E-mail addresses: kjlz@ahu.edu.cn (Z. Lv), wxp2001@ahu.edu.cn (X. Wu).

https://doi.org/10.1016/j.neunet.2023.09.041
Received 5 June 2023; Received in revised form 18 August 2023; Accepted 24 September 2023
Available online 25 September 2023
0893-6080/

operations (Abdelaziz, Gong, & Stylianou, 2021; Luo & Mesgarani, 2019; Pandey & Wang, 2019; Zhang, Chen, et al., 2022). However, the magnitude and phase are only implicitly handled in the optimization process of the waveform samples, they lack explicit modeling, and thus minimizing the distortion in the time domain cannot guarantee the estimation accuracy of speech spectrum. In contrast, complex-domain based methods (Hu, Liu, et al., 2020; Tan & Wang, 2020; Zhang, Gao, & Liu, 2022) convert the joint optimization toward the magnitude and phase into that of the real and imaginary (RI) components, which exhibit similar spectral structure to the spectral magnitude. Therefore, both the magnitude and phase can be effectively recovered, which partly explains the superiority of the complex-domain methods over magnitude-domain methods. Note that both magnitude-domain and complex-domain methods belong to the T–F domain and here we distinguish between them only for better illustrations.

Studies using time-domain or complex-domain methods consider phase information to some extent, but the magnitude of the speech tends to compensate for inaccurate phase estimation (Wang, Wichern, & Le Roux, 2021), resulting in limited enhancement performance. For this reason, multi-stage strategies are proposed to investigate the performance improvement from progressive or multi-domain perspective. Complicated tasks are often decomposed into smaller and more achievable goals, and are then optimized sequentially to achieve the final goal. For example, in Li, Yuan, Zheng, and Li (2020), researchers used the SNR as a progressive criterion for multi-stage training to gradually improve speech quality and enhance the enhancement effect. In Wang and Wang (2022), a cascade method was proposed that combines the magnitude domain, complex domain, and time domain for speech enhancement, which achieved excellent enhancement performance. It shows that the combination of different processing domains can yield better performance than single-domain. However, such a naïve cascade design may lack the mechanism analysis of why the assembling of multi-domain can provide better performance.

To overcome these limitations, we propose CompNet, a single-channel speech enhancement framework. Unlike previous multi-stage networks, CompNet achieves its ultimate goal from a complementary perspective. CompNet consists of two components: a time-domain pre-processing module and a T–F domain post-processing module, which are trained in an end-to-end manner. The pre-processing module uses the time-domain network to enhance the speech waveform and optimize the sampling points, but the magnitude and phase at this stage may not reach their optimal solution. To solve this problem, we introduce a post-processing module in the T–F domain. This module is devised as a parallel branching structure to optimize magnitude and phase independently (Li, Liu, et al., 2021; Yin, Luo, Xiong, et al., 2020), aiming to obtain the better solution as much as possible. As such, the proposed CompNet can achieve the feasible enhancement performance with the help of the complementary nature between the two modules.

The contributions of this paper are two-fold. First, different from previous works with naive cascading multiple networks without rationality illustrations, this paper attempts to provide a different perspective to understand the improvements caused by multi-domain processing. Specifically, in the time-domain, despite the waveform samples can be feasibly recovered, the distribution of the T–F bins may still be partly different from the target spectrum. Therefore, we specially devise a dual-path network as the post-processing module to further push the estimated spectrum to get close to the target as much as possible in the frequency domain. Second, we follow the "from-coarse-to-fine" concept for pipeline design. After the processing by the first stage, some residual noise can still exist and thus we adopt the second network for further noise suppression. We conduct experiments on the WSJ0-SI84 dataset and the VoiceBank + Demand dataset to evaluate the proposed approach. The experimental results demonstrate the highly competitive nature of CompNet compared to existing systems.

This article is structured as follows: Section 2 presents the problem formulation, followed by a detailed description of the proposed framework in Section 3. The experimental setup is outlined in Section 4, while Section 5 covers the results and analysis. Section 6 gives the discussions. Finally, conclusions are presented in Section 7.

2. Problem formulation

In the context of monophonic speech enhancement, we define the noisy speech signal $y[t]$ in the time domain as a combination of clean target speech $s[t]$ and background noise $n[t]$,

$$y[t] = s[t] + n[t] \tag{1}$$

where t represents a specific time sample. By applying the short-time Fourier transform (STFT) function, we can convert the noisy speech signal $y[t]$ from the time domain to the T–F domain,

$$Y(l, f) = S(l, f) + N(l, f) \tag{2}$$

where $Y(l, f)$, $S(l, f)$, and $N(l, f)$ denote the mixed speech signal, the clean speech signal, and the background noise signal, respectively. In the T–F domain, l is the index of the time frame, and f is the index of the frequency bin. For notation convenience, (l, f) is omitted in the rest of the paper.

There are two distinct forms of speech representation following STFT. In the polar coordinates, it can be expressed as a combination of magnitude and phase,

$$|Y|e^{i\theta_Y} = |S|e^{i\theta_S} + |N|e^{i\theta_N} \tag{3}$$

where $|\cdot|$ denotes magnitude and θ denotes phase. In the Cartesian coordinate, the above representation can be formulated as the addition of the real component and the imaginary component,

$$Y = Y_r + jY_i = \left(S_r + N_r\right) \quad +j\left(S_i + N_i\right) \tag{4}$$

where the subscripts r and i denote the real and imaginary parts, respectively, and j is the imaginary unit.

3. Complementary framework

In this section, we will delve into the details of the proposed CompNet. However, before we do so, it is crucial to introduce two key components of the framework: the squeezed time convolution module (S-TCM) and the Nested U-Structure. These components play a pivotal role in the overall architecture and warrant a thorough explanation before moving on to the introduction of CompNet.

3.1. Squeezed temporal convolutional module

Maintaining long-term memory is an important aspect of neural network structures. To address this challenge, researchers proposed the long short-term memory (LSTM) network (Hochreiter & Schmidhuber, 1997), which allows neurons to retain contextual memory in their pipelines. Subsequently, the temporal convolutional module (TCM) introduced in Bai, Kolter, and Koltun (2018) was found to be more effective in time series modeling than LSTM, and is currently widely applied in speech separation (Fan et al., 2020; Luo & Mesgarani, 2019). The structure of TCM is shown in Fig. 1(a), which includes a 1×1 input convolution to increase channel dimensionality, a 1×1 output convolution to restore the number of channels, and a depth-dilated convolution (DD-Conv) to capture long-term temporal dependencies. Additionally, TCM incorporates parameterized ReLU (PReLU) (Guimarães, Nagano, & Silva, 2020) and normalization layers between adjacent convolutions and utilizes residual connections to facilitate information flow.

Recently, a compressed version of TCM called squeezed temporal convolutional module (S-TCM) was proposed in Li, Liu, Zheng, Fan, and Li (2021), as shown in Fig. 1(b). S-TCM differs from TCM mainly in two aspects. First, in order to improve long-range modeling, the DD-Conv is replaced by a traditional dilated convolution (D-Conv). Second, a gating branch is added, which is similar to the main branch except for an additional sigmoid activation function controlling the output within the range of (0, 1). Experiments in Li, Liu, Zheng, et al. (2021) showed that S-TCM has about 72% fewer parameters than TCM, while still providing comparable performance. Therefore, we use the compressed version of TCM, as the default choice in our experiments.

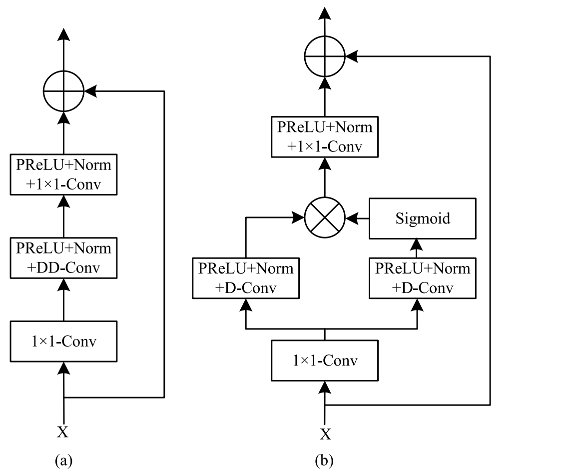

Fig. 1. (a) Structure illustration of the TCM. (b) Structure illustration of the Squeezed TCM (S-TCM).

Table 1
Network details of TCNN.

Layer name	Input size	Hyperparameters	Output size
Convglu 1	1 × T × 320	(2, 3), (1, 2), 64	64 × T × 159
Convglu 2	64 × T × 159	(2, 3), (1, 2), 64	64 × T × 79
Convglu 3	64 × T × 79	(2, 3), (1, 2), 64	64 × T × 39
Convglu 4	64 × T × 79	(2, 3), (1, 2), 64	64 × T × 19
Convglu 5	64 × T × 19	(2, 3), (1, 2), 64	64 × T × 9
Convglu 6	64 × T × 9	(2, 3), (1, 2), 64	64 × T × 4
Reshape_1	64 × T × 4	–	256 × T
S-TCMs	256 × T	[1, 2, 5, 9] × 3	256 × T
Reshape_2	256 × T	–	64 × T × 4
Deconvglu 6	128 × T × 4	(2, 3), (1, 2), 64	64 × T × 9
Deconvglu 5	128 × T × 9	(2, 3), (1, 2), 64	64 × T × 19
Deconvglu 4	128 × T × 19	(2, 3), (1, 2), 64	64 × T × 39
Deconvglu 3	128 × T × 39	(2, 3), (1, 2), 64	64 × T × 79
Deconvglu 2	128 × T × 79	(2, 3), (1, 2), 64	64 × T × 159
Deconvglu 1	128 × T × 159	(2, 3), (1, 2), 64	64 × T × 320

3.2. Nested U-structure

The U-Net model was first proposed in Olaf et al. (2015) and is now widely used in the field of speech processing (Choi et al., 2021; Guimarães et al., 2020; Xu, Xu, Kong, & Xu, 2022). The U-Net consists of a contracting path and an expansive path. Each layer of the contracting path contains a downsampling convolution, a normalization layer, and a PReLU activation function. Similarly, each layer in the expansive path consists of an upsampling convolution, followed by a normalization layer and a PReLU activation function. In U-Net, skip connections are used to connect the feature maps of each layer in the contracting path to the feature maps of the same size in the expansive path, facilitating the flow of information through the network.

Recently, a new version of U-Net called U^2-Net was introduced in Qin and Zhang (2020). U^2-Net adopts a two-layer nested U-shaped structure, with a similar overall framework as U-Net. In contrast, the U^2-Net features residual u-blocks (RSUs) with different receptive field sizes at each stage. This architecture allows for capturing more contextual information at different scales, thereby endowing U^2-Net with rich multi-scale features. Therefore, in this study, we use the U^2-Net architecture for experiments to achieve better performance.

3.3. The proposed framework

The design of the speech enhancement framework is depicted in Fig. 2. The CompNet consists of two fundamental modules: a time domain pre-processing module and a T–F domain post-processing module. The original noisy speech first passes the time-domain preprocessor for initial estimation. The T–F domain post-processing module then filters the magnitude and refines the phase. In the following subsections, the time domain pre-processing module and the T–F domain post-processing module are described in detail.

3.3.1. Preprocessing module in the time domain

We use a temporal convolutional neural network (TCNN) structure as the time domain preprocessing module, whose input is the original noisy speech frames,

$$\tilde{s}_{time} = TCNN\left(y; \Phi_1\right) \tag{5}$$

where $TCNN\left(\bullet; \Phi_1\right)$ denotes the time-domain network function and $\tilde{s}_{time}$ denotes the speech processed by the time-domain network. TCNN consists of three components: the encoder, the decoder, and the S-TCMs. The TCNN module is portrayed in Fig. 2. The encoder–decoder structures are symmetric, each of which using a stack of five 2D gated convolutional (2D-ConvGLU) layers. Both the encoder and decoder have the same step size and convolution kernel, and the difference is that the encoder uses convolution to downsample along the frequency axis whereas the decoder uses transposed convolution. Each convolution is followed by Layer Normalization (LN) and PReLU. TCNN uses S-TCMs in the bottleneck to model the time dependency. The skip connection is adopted between the encoder and decoder.

Before feeding into the T–F domain network, we also need to combine the preprocessed magnitude $|\tilde{S}_{time}|$ with the original phase θ_Y as $\tilde{S}_{mag}$ and combine the preprocessed phase $\theta_{\tilde{S}_{time}}$ with the original magnitude $|Y|$ as $\tilde{S}_{phase}$,

$$\tilde{S}_{phase} = |Y|cos(\theta_{\tilde{S}_{time}}) + j|Y|sin(\theta_{\tilde{S}_{time}}) \tag{6}$$

$$\tilde{S}_{mag} = |\tilde{S}_{time}|cos(\theta_Y) + j|\tilde{S}_{time}|sin(\theta_Y) \tag{7}$$

where $\tilde{S}_{phase}$ and $\tilde{S}_{mag}$ refer to the complex spectrum after preliminary enhancement of the phase and the complex spectrum after magnitude enhancement, respectively. While $\tilde{S}_{mag}$ estimates the magnitude but requires phase correction, $\tilde{S}_{phase}$ estimates the phase but needs magnitude refinement. Thus, these two components form a complementary pair. The complementary pair encompasses both the information estimated by the time-domain model and the original noisy information, thereby jointly participating in the processing of the second stage.

3.3.2. Postprocessing module in the T-F domain

It is worth noting that the T–F domain post-processing module is not a symmetric structure, consisting mainly of the encoder part of the U^2-Net (U^2-Encoder), the Gain Branch, the Resi Branch, and the decoder. The inputs of the T–F domain post-processing module are $\tilde{S}_{phase}$ and $\tilde{S}_{mag}$, and its main objective in the T–F domain is magnitude filtering and phase refining, as shown in Fig. 2. The U^2-Encoder contains four RSUs and a gated convolutional layer, which we use to extract the features,

$$En_x = Encoder_{U^2}(Cat(\tilde{S}_{phase}, \tilde{S}_{mag})) \tag{8}$$

where $Cat(\bullet)$ denotes the concatenation function and $Encoder_{U^2}(\bullet)$ is the encoder function with the U^2-Net structure. The output obtained through the encoder is denoted as En_x. The Gain Branch is responsible for estimating the gain to facilitate magnitude filtering, whereas the Resi Branch is responsible for estimating the residual component for phase refining,

$$gain = GainBranch\left(En_x, \tilde{S}_{mag}; \Phi_2\right) \tag{9}$$

$$resi = ResiBranch\left(En_x, \tilde{S}_{phase}; \Phi_3\right) \tag{10}$$

where $GainBranch(\bullet)$ refers to the Gain Branch function and $ResiBranch$ $(\bullet)$ refers to the Resi Branch function. $GainBranch(\bullet)$ and $ResiBranch(\bullet)$

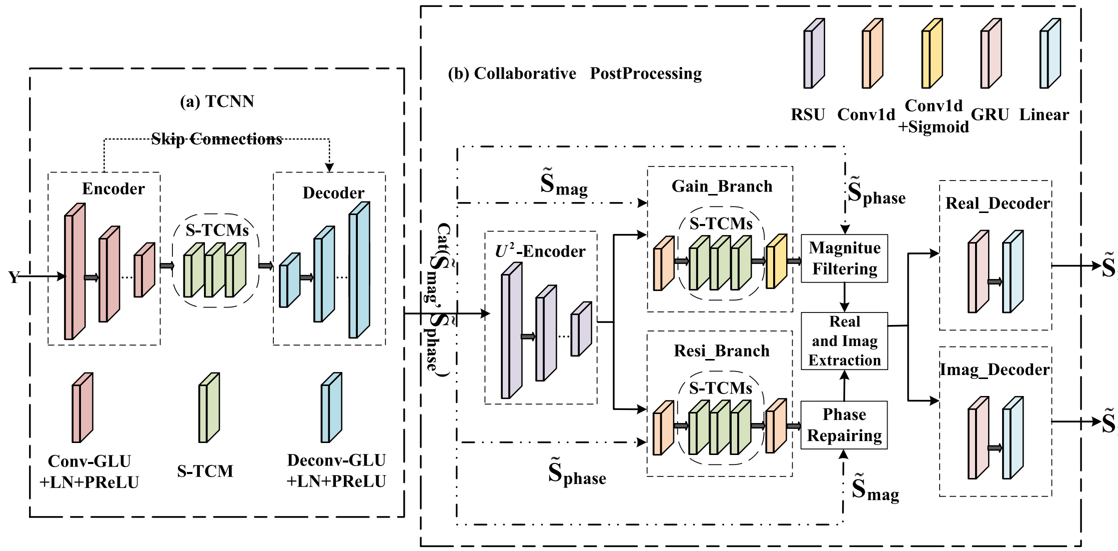

Fig. 2. An illustration of CompNet, which consists of two sub-modules, the temporal convolutional neural network (TCNN) and collaborative post-processing. The TCNN performs a global coarse estimation of the target speech in the time domain. The collaborative post-processing complements the spectral information by predicting the phase residuals and magnitude gain, resulting in clearer speech. (a): The TCNN consists of an encoder, decoder and stacked S-TCMs. (b): The U^2-Encoder, Gain Branch, Resi Branch and Decoder are part of the collaborative post-processing. (The encoder and decoder modules use 2D convolutional layers in GLU format, and the corresponding blocks are called Conv-GLU and Deconv-GLU blocks, respectively.).

all use the S-TCMs structure to model the time dependence. The obtained $gain$ and $resi$ are subsequently utilized for collaborative restoration of both $\tilde{S}_1$ and $\tilde{S}_2$,

$$\tilde{S}_1 = gain * \tilde{S}_{phase} \tag{11}$$

$$\tilde{S}_2 = resi + \tilde{S}_{mag} \tag{12}$$

We extract the real and imaginary parts of $\tilde{S}_1$ and $\tilde{S}_2$ separately. Then, combine the real parts of $\tilde{S}_1$ and $\tilde{S}_2$ to derive the $real_x$, and the imaginary parts of $\tilde{S}_1$ and $\tilde{S}_2$ to derive the $imag_x$. Finally, $real_x$ and $imag_x$ are fed into a decoder consisting of linear layers, respectively,

$$real_x = Decoder_{real}(GRU(real_x)) \tag{13}$$

$$imag_x = Decoder_{imag}(GRU(imag_x)) \tag{14}$$

where $GRU(\bullet)$ is a recurrent neural network (RNN) function (Chung et al., 2014) that we utilize to process the speech information before passing it into the decoder. $Decode_{real}(\bullet)$ serves as the decoder for the real part, while $Decode_{imag}(\bullet)$ functions as the decoder for the imaginary part. Finally, we combine the $real_x$ and $imag_x$ to obtain the post-processed speech signal in the T–F domain.

Tables 1 and 2 summarize the network design details for each module. Layer names indicate the function and location of the corresponding layer or block. The input is specified with (Channels ×TimeSteps ×FreqChannels) for 3D-format and (FreqChannels ×TimeSteps) for 2D-format. For the U^2-Encoder, the hyperparameters are specified with (KernelSize1, KernelSize2, Stride, T, Channels) format. Here, KernelSize1 and KernelSize2 respectively denote the kernel size of 2D-ConvGLU and the convolutional kernel size within the U-Net block. For one-dimensional convolutions (1D-conv), the hyperparameters are specified as (KernelSize, Channel). In GRU and Linear layers, the hyperparameters only include the number of output channels. In S-TCMs, the hyperparameters are specified in [DilationRate] format and the kernel size for dilation convolution is 5.

3.4. Loss function

The training objective of our CompNet consists of two parts, corresponding to the outputs generated by the two modules. In our time-domain preprocessing module, we train the time-domain network with SNR loss until convergence,

$$\begin{cases} \mathbf{s}_{\text{target}} & := \frac{\langle \hat{\mathbf{s}}, \mathbf{s}\rangle \mathbf{s}}{\|\mathbf{s}\|^2} \\ \mathbf{e}_{\text{noise}} & := \hat{\mathbf{s}} - \mathbf{s}_{\text{target}} \\ \text{SNR} & := 10\log_{10}\frac{\|\mathbf{e}_{\text{noise}}\|^2}{\|\mathbf{s}_{\text{target}}\|^2} \end{cases} \tag{15}$$

where $\hat{s}$ denotes the predicted speech, s_{target} is the target clean speech, $\|\mathbf{s}\|^2 = \langle \mathbf{s}, \mathbf{s}\rangle$ denotes the signal power.

In the T–F domain post-processing module, we explicitly separate the magnitude and phase. Therefore, the T–F domain loss is defined as:

$$\begin{cases} \mathcal{L}_{RI} = \left\|\tilde{S}_r - S_r\right\|_F^2 + \left\|\tilde{S}_i - S_i\right\|_F^2 \\ \mathcal{L}_{Mag} = \left\|\sqrt{\left|\tilde{S}_r\right|^2 + \left|\tilde{S}_i\right|^2} - \sqrt{\left|S_r\right|^2 + \left|S_i\right|^2}\right\|_F^2 \\ \mathcal{L}_{T-F} = \alpha\mathcal{L}_{Mag} + (1-\alpha)\mathcal{L}_{RI} \end{cases} \tag{16}$$

In our CompNet, the time domain preprocessing module and the T–F domain post-processing module are trained jointly. This training method helps to back-propagate the gradient from the T–F domain post-processing module to the time domain preprocessing module. Thus, the total loss can be expressed as:

$$\mathcal{L}_{total} = \gamma_1 SNR + \gamma_2(\alpha\mathcal{L}_{Mag} + (1-\alpha)\mathcal{L}_{RI}) \tag{17}$$

where parameter α is set to 0.5 by default, parameter γ_1 is set to 0.2 and parameter γ_2 is specified as 1.

4. Dataset and experimental setup

4.1. Dataset preparation

To evaluate the performance of CompNet, we conducted extensive experiments using two datasets: WSJ0-SI84 (Paul & Baker, 1992) and VoiceBank + Demand Valentini-Botinhao, Wang, Takaki, and Yamagishi (2016). WSJ0-SI84 consists of 7138 clear speech samples from 83 speakers, with 41 female and 42 male speakers. We selected 77 speakers, and used 5428 and 957 clear speech samples for training and validation, respectively. For testing, we used two sets of 150 speech samples each, from six speakers (three male and three female). In the

Table 2
Network details of collaborative postprocessing.

U^2-encoder			
Layer name	Input size	Hyperparameters	Output size
RSU 1	4 × T × 161	64, (2, 5), (2, 3), (1, 2)	64 × T × 79
RSU 2	64 × T × 79	64, (2, 5), (2, 3), (1, 2)	64 × T × 39
RSU 3	64 × T × 39	64, (2, 5), (2, 3), (1, 2)	64 × T × 19
RSU 4	64 × T × 19	64, (2, 5), (2, 3), (1, 2)	64 × T × 9
Last_convglu	64 × T × 9	64, (2, 5), (1, 2)	64 × T × 4

Gain_Branch				Resi_Branch			
Layer name	Input size	Hyperparameters	Output size	Layer name	Input size	Hyperparameters	Output size
Reshape	64 × T × 4	–	417 × T	Reshape	64 × T × 4	–	518 × T
Conv1d	417 × T	1, 256	256 × T	Conv1d	518 × T	1, 256	256 × T
S-TCMs	256 × T	[1, 2, 5, 9] × 3	256 × T	S-TCMs	256 × T	[1, 2, 5, 9] × 3	256 × T
Conv1d	256 × T	1, 161	161 × T	Conv1d	256 × T	1, 322	322 × T
Magnitude filtering	161 × T	–	2 × T × 161	Phase Refining	322 × T	–	2 × 161 × T

Layer name	Input size	Hyperparameters	Output size
Real extraction	(2 × T × 161),(2 × T × 161)	–	322 × T
Imag extraction	(2 × T × 161), (2 × T × 161)	–	322 × T

Real_Decoder				Imag_Decoder			
Layer name	Input size	Hyperparameters	Output size	Layer name	Input size	Hyperparameters	Output size
GRU	322 × T	161	161 × T	GRU	322 × T	161	161 × T
Linear	161 × T	161	161 × T	Linear	161 × T	161	161 × T

first set, the speaker's speech data was included in the training set, and we referred to these speakers as visible speakers. In the second set, the speaker's information was not included in the training set, and we called these speakers invisible speakers. The purpose of using invisible speakers was to test the model's ability to generalize to different speakers. We created 150,000 training samples and 10,000 validation samples by randomly selecting 20,000 types of noise from the DNS challenge noise set (about 55 h in duration) and using them to mix noisy speech at SNRs ranging from −5 dB to 0 dB (about 300 h in total). Two different test datasets, called test set 1 and test set 2, were used to evaluate our method. Test set 1 was exclusively designed for conducting ablation experiments. To achieve this, we selected two challenging noises from the NOISEX92 database (Varga & Steeneken, 1993) (babble and factory1) and mixed them with clear speech at five different SNRs: −6 dB, −3 dB, 0 dB, 3 dB, and 6 dB. Each case contained 150 pairs of speech samples. Test set 2 was utilized to compare the performance of our proposed method against the baseline. To achieve this comparison, we randomly selected noise from the NOISEX92 database (Varga & Steeneken, 1993) and mixed it with clean speech at randomly selected SNRs (−6 dB, −3 dB, 0 dB, 3 dB, or 6 dB). Both the seen and unseen test sets 2 consisted of 3558 noisy-clean pairs, respectively.

VoiceBank + Demand is a dataset containing 30 speakers. Of these, 28 speakers were used in the training set and the remaining 2 speakers were used in the test dataset. The training dataset consists of 11,572 noisy speech samples, obtained by mixing clean speech with 40 different types of noise at SNRs of {0 dB, 5 dB, 15 dB, 20 dB}. The test simulation dataset consists of 824 noisy speech samples, obtained by mixing clean speech with 20 different types of noise at SNRs of {2.5 dB, 7.5 dB, 12.5 dB, 17.5 dB}.

4.2. Experimental setup

To ensure stability during training, all speech recordings were sampled at a frequency of 16 kHz. Recordings longer than 8 s were truncated to 8 s, while recordings shorter than 8 s were padded with zeros. We extracted frames using a rectangular window size of 20 ms and an overlap time of 10 ms. For the purpose of STFT operations, we utilized a 20 ms Hamming window with a 50% overlap between adjacent time frames. This resulted in the utilization of a 320-point STFT, which produced a 161-dimensional spectrum. Recently, power compression techniques have been shown to improve performance (Li, Zheng, et al., 2021), and we apply this strategy to the input and target, i.e. $|Y|^{\beta} e^{j\theta_Y}, |S|^{\beta} e^{j\theta_S}$, where β= 0.5 is considered empirically optimal.

During the training process, we established a batch size of 8 and utilized the Adam optimizer (Kingma & Ba, 2015) for stochastic gradient descent optimization with β_1 set to 0.9 and β_2 set to 0.999. To prevent an asymptotic explosion, we applied asymptotic clipping with a maximum value of 5.0. The model was trained for 60 periods with an initial learning rate of 0.0005. If the validation loss did not decrease for two consecutive epochs, the learning rate was halved. If the validation loss did not decrease for three consecutive epochs, the network would be stopped early.

4.3. Baseline models

This work conducts experiments on two different datasets, and we do not use the same baselines for our experiments. In the experiments based on the WSJ0-SI84 dataset, we selected eight baselines. Four of these baselines are intended for causal systems, which include LSTM (Chen, Wang, oho, Wang, & Healy, 2016), convolutional recurrent neural network (CRN) (Tan & Wang, 2018), gated convolutional recurrent networks (GCRN) (Tan & Wang, 2020), and ConvTasNet (Luo & Mesgarani, 2019). The LSTM and CRN models operate in the magnitude domain. The CRN model has an encoder–decoder structure with an intermediate LSTM for timing modeling. The GCRN method is an improved version of CRN that replaces regular convolutions in the encoder and decoder with their GLU versions. ConvTasNet is a time domain model, an advanced end-to-end speaker separation system that predicts waveform samples directly rather than using STFT for conversion. It has excellent performance, even for speech enhancement tasks. The other four baselines are for non-causal systems such as BLSTM, BCRN, BGCRN, and the non-causal ConvTasNet. BLSTM, BCRN, and BGCRN are similar to LSTM, CRN, and GCRN, but with bidirectional versions of LSTM replacing all LSTM layers. Noncausal ConvTasNet allows the use of future information. Each individual baseline model operates with the default parameter settings as prescribed in this research paper. Note that ConvTasNet originally used an 8 kHz sampling rate, but was extended to 16 kHz for speech enhancement. All models were trained on a GPU with a Pytorch platform (Paszke et al., 2017).

In the context of experiments conducted on the VoiceBank + Demand dataset, twelve different baseline models were selected for comparative analysis. The first group of models focused on speech enhancement through the use of generative adversarial techniques in the T–F

Table 3
Regarding the ablative study of different post-processing types and encoders in post-processing networks. “vanilla” represents the simultaneous processing of magnitude and phase, while “collaborative” represents the separate processing of magnitude and phase. “u” represents the use of U-Net architecture as the encoder in the post-processing network. “u^2” represents the use of U^2-Net architecture as the encoder in the post-processing network.

(a) Seen speaker test set 1

	Metrics	PESQ						ESTOI (%)						SDR (dB)					
	SNR (dB)	−6	−3	0	3	6	Avg	−6	−3	0	3	6	Avg	−6	−3	0	3	6	Avg
Babble noise	Noisy	1.55	1.71	1.87	2.06	2.25	1.89	26.66	34.43	42.81	52.47	61.85	43.64	−5.89	−2.93	0.04	3.04	6.03	0.06
	vanilla + u	1.90	2.24	2.51	2.78	2.99	2.48	54.34	65.00	72.84	79.17	83.70	71.01	4.26	7.04	9.37	11.23	12.70	8.50
	vanilla + u^2	1.96	2.33	2.59	2.85	3.05	2.56	57.53	67.84	74.99	80.77	84.88	73.20	4.63	7.43	9.64	11.41	12.83	9.19
	collaborative + u	1.97	2.34	2.62	2.88	3.08	2.58	58.94	68.92	76.01	81.47	85.52	74.17	4.97	7.60	9.85	11.55	13.00	9.40
	collaborative + u^2	**1.99**	**2.35**	**2.65**	**2.91**	**3.11**	**2.60**	**60.05**	**70.00**	**77.00**	**82.18**	**86.08**	**75.06**	**5.15**	**7.83**	**10.02**	**11.69**	**13.09**	**9.56**
Factory1 noise	Noisy	1.43	1.60	1.79	1.98	2.20	1.80	26.12	34.66	44.10	54.04	64.21	44.62	−5.90	−2.93	0.05	3.04	6.03	0.06
	vanilla + u	2.07	2.40	2.65	2.85	3.04	2.60	52.98	64.60	72.27	78.50	83.13	70.29	5.98	8.28	10.06	11.54	12.81	9.73
	vanilla + u^2	2.19	2.51	2.75	2.94	3.11	2.70	56.34	67.42	74.54	80.28	84.45	72.61	6.46	8.67	10.36	11.74	13.01	10.05
	collaborative + u	2.17	2.51	2.76	2.96	3.13	2.71	57.35	68.27	75.38	80.93	84.92	73.37	6.31	8.55	10.30	11.73	12.97	9.97
	collaborative + u^2	**2.21**	**2.54**	**2.78**	**2.98**	**3.16**	**2.73**	**59.20**	**69.52**	**76.26**	**81.80**	**85.57**	**74.47**	**6.52**	**8.75**	**10.46**	**11.88**	**13.09**	**10.14**

(b) Unseen speaker test set 1

	Metrics	PESQ						ESTOI(%)						SDR (dB)					
	SNR (dB)	−6	−3	0	3	6	Avg	−6	−3	0	3	6	Avg	−6	−3	0	3	6	Avg
Babble noise	Noisy	1.45	1.63	1.82	2.02	2.22	1.83	24.32	31.74	39.79	48.50	57.37	40.34	−5.90	−2.94	0.04	3.03	5.78	−0.01
	vanilla + u	1.84	2.20	2.47	2.70	2.90	2.42	51.21	62.07	70.89	77.28	82.00	68.69	4.15	7.11	9.62	11.42	12.92	9.04
	vanilla + u^2	1.93	2.27	2.55	2.77	2.98	2.50	54.60	64.97	72.96	79.24	83.53	71.06	4.63	7.44	9.87	11.64	13.09	9.33
	collaborative + u	1.92	2.28	2.56	2.79	2.98	2.51	55.19	65.83	73.88	79.76	83.73	71.68	4.81	7.54	10.01	11.75	13.18	9.46
	collaborative + u^2	**1.92**	**2.29**	**2.59**	**2.82**	**3.01**	**2.53**	**56.60**	**67.00**	**74.74**	**80.58**	**84.46**	**72.68**	**5.10**	**7.82**	**10.22**	**11.96**	**13.32**	**9.69**
Factory1 noise	Noisy	1.38	1.55	1.75	1.96	2.18	1.76	23.88	31.78	40.77	50.49	60.15	41.41	−5.89	−2.94	0.04	3.04	6.03	0.05
	vanilla + u	2.04	2.34	2.60	2.79	2.94	2.54	49.36	61.33	70.79	77.00	81.72	68.04	6.16	8.52	10.39	11.84	13.08	10.00
	vanilla + u^2	2.13	2.45	2.69	2.86	3.01	2.63	53.12	64.62	73.31	78.79	83.14	70.60	6.71	8.98	10.73	12.11	13.31	10.37
	collaborative + u	2.12	2.44	2.69	2.87	3.01	2.63	53.50	65.09	73.73	79.27	83.40	71.00	6.46	8.76	10.59	12.05	13.26	10.22
	collaborative + u^2	**2.15**	**2.47**	**2.71**	**2.88**	**3.02**	**2.65**	**55.27**	**66.48**	**74.59**	**79.91**	**84.00**	**72.05**	**6.70**	**8.97**	**10.76**	**12.20**	**13.38**	**10.40**

domain, such as MMSE-GAN (Chen et al., 2016), MetricGAN (Fu, Liao, Tsao, & Lin, 2019), and SEGAN (Pascual, Bonafonte, & Serrà, 2017). The second group comprised methods that operate on the time domain for speech enhancement, such as SE-Flow−μ (Strauss & Edler, 2021), SEAMNET (Borgström & Brandstein, 2021), Wave-U-Net (Macartney & Weyde, 2018), and CPTNN (Wang, He, & Zhu, 2022). The third group included baseline models that operate in the complex domain, specifically DCCRN (Hu, Liu, et al., 2020), S-DCCRN (Lv et al., 2022), and DCTCN (Zou & Zhu, 2021). In addition, there are cross-domain networks like TFT-Net (Tang, Luo, Zhao, Xie, & Zeng, 2020) and parallel network DBT-Net (Yu et al., 2022).

4.4. Evaluation metrics

Depending on the dataset, we have used different metrics in this paper. The ablation experiments were specifically evaluated on the WSJ0-SI84 dataset (Paul & Baker, 1992) using perceptual evaluation of speech quality (PESQ) (Rix, Beerends, Hollier, & Hekstra, 2001), extended short-term objective intelligibility (ESTOI) (Jensen & Taal, 2016), and signal-distortion ratio (SDR) (Vincent, Sawada, Bofill, Makino, & Rosca, 2007). For comparison with other baselines, the SDR was replaced by DNS-MOS (Naderi & Cutler, 2021), a recently proposed metric that follows ITU-T Rec.P.835 and is highly correlated with subjective human ratings. The PESQ provides a measure of speech quality ranging from −0.5 to 4.5, while the ESTOI measures speech intelligibility as a percentage between 0 and 1. The SDR is widely recognized for blind source separation and is a valid indicator of the degree of speech distortion.

On the other hand, for the VoiceBank + Demand dataset, we used wideband PESQ (Recommendation ITU-T P ITU, 2007) and three additional target metrics associated with MOS (Hu & Loizou, 2007), namely CSIG, CBAK, and COVL, to assess speech quality. For all metrics used in our experiments, higher values indicate better speech quality.

5. Results and analysis

5.1. Ablation study

To ensure experimental integrity, the ablation study of the model was divided into two sub-experiments, each of which was based on the WSJ0-SI84 test set 1. The first sub-experiment compares CompNet networks with different functional modules to demonstrate that the proposed CompNet has the best overall performance. In the second sub-experiment, the performance of the proposed CompNet model was evaluated at different stages to confirm its validity. The results of these two sub-experiments are presented in Table 3 and Fig. 3, respectively. All models trained for the ablation experiments were causal systems.

We have conducted evaluations of various versions of CompNet, which are detailed in Table 3. The “vanilla” version refers to the simultaneous processing of magnitude and phase in the T–F domain network. The “collaborative” version is divided into two branches within the T–F domain module, allowing for independent processing of magnitude and phase. The letter “u” indicates the use of U-Net as the encoder in the T–F domain module, while “u^2” indicates the use of U^2-Net as the encoder. Our first variant uses the vanilla strategy and the U-Net encoder and is marked “vanilla + u”. The remaining variants are denoted in order as “vanilla + u^2”, “collaborative + u”, and “collaborative + u^2” respectively. The following conclusions can be drawn from ablation experiment I.

- The variants show consistent performance across different test sets and different SNRs. Comparing “vanilla + u” with “vanilla + u^2” or “collaborative + u” with “collaborative + u^2”, we can confirm that the U^2-Encoder outperforms the U-Encoder.
- We can observe that the collaborative strategy performs better than the vanilla strategy when comparing “vanilla + u” with “collaborative + u” or “vanilla + u^2” with “collaborative + u^2”. Furthermore, the improvement of “collaborative + u^2” over “vanilla + u” is mainly since the collaborative strategy and the U^2-Encoder can achieve a “one-plus-one is greater than two” effect.

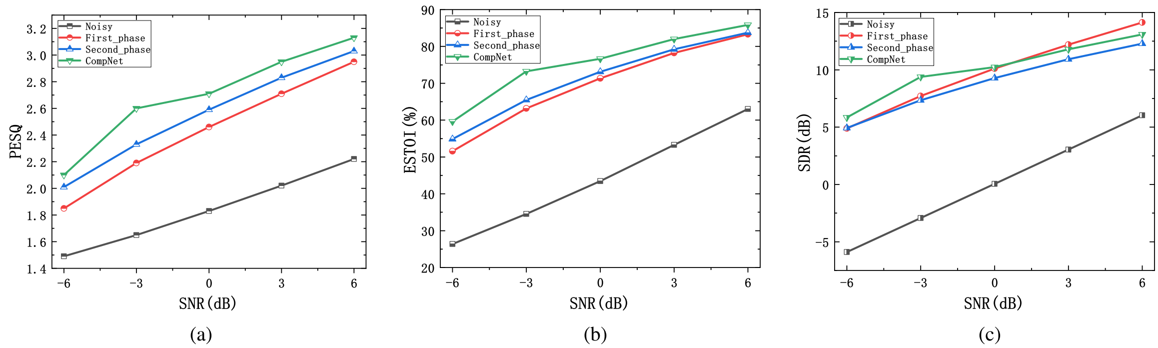

Fig. 3. PESQ, ESTOI, and SDR under different noisy conditions for different models. The values in each input SNR are averaged from all the seen test set 1. The input SNR value ranges from −6 dB to 6 dB with an interval of 3 dB.

Table 4
Number of trainable parameters and MACs for different enhancement models.

Model	Parameters	MACs
LSTM	29.04 M	2.94 G/s
CRN	17.58 M	2.44 G/s
GCRN	9.77 M	2.42 G/s
ConvTasNet	5.00 M	5.23 G/s
CompNet	4.26 M	5.92 G/s

In ablative experiment one, we found that the results were not affected by the noise or test dataset. Therefore, in ablative experiment two, we only enhanced the visible mixed speech. Ablative experiment two investigated the functions of CompNet at different stages. From Fig. 3, we can obtain the following observations. "First_phase" refers to using TCNN to enhance noisy speech, while "Second_phase" means filtering the magnitude and correcting the phase of the noisy speech using a T–F domain network.

- Firstly, CompNet consistently outperforms both the first and second stages in terms of PESQ and ESTOI, highlighting the importance of complementarity. Specifically, compared to the second stage, the speech enhanced by CompNet exhibits better optimization at the sample level, while compared to the first stage, CompNet can supplement the details of the speech.
- Secondly, time-domain networks tend to perform better on SDR, but CompNet outperforms the First_phase on SDR at low SNRs and performs similarly to it at high SNRs. This indicates that CompNet has a very stable and effective speech enhancement capability at low SNRs.

5.2. Model complexity comparison

For the causal speech enhancement system, we need to consider several statistics in a realistic scenario, such as the model size, and the number of multiplicative accumulation operations (MACs) per second. Table 4 lists the four causal baselines as well as the data sizes of the proposed CompNet. It is important to note that input samples are all set to one second of audio, so our experiments are fair. According to Table 4, CompNet has smaller model parameters but is relatively more computationally intensive.

5.3. Comparisons with baselines on WSJ0-SI84 corpus

On the WSJ0-SI84 test set 2, we evaluated the proposed CompNet against some classical baseline models. The test set 2 is divided into a visible speaker dataset and an unseen speaker dataset. Evaluation indicators: PESQ, ESTOI, MOS_OVLR. We have a total of eight baselines, four of which are causal systems and the rest belong to non-causal systems.

Table 5
Estimation and comparison of PESQ, ESTOI (%), and overall DNS-MOS (MOS_OVLR) metrics among different enhancement models in the seen test set 2.

System	Causal	PESQ	ESTOI (%)	MOS_OVLR
Noisy		1.99	52.82	1.84
LSTM	✓	2.71	73.29	2.74
CRN	✓	2.77	74.16	2.75
GCRN	✓	2.89	78.70	2.83
ConvTasNet	✓	2.76	78.37	2.74
CompNet	✓	**2.95**	**79.72**	**2.91**
BLSTM	×	2.93	78.81	2.97
BCRN	×	2.95	78.80	2.96
BGCRN	×	3.06	81.29	2.97
ConvTasNet_noncau	×	3.06	**83.73**	**3.11**
CompNet_noncau	×	**3.14**	83.23	3.04

Table 6
Estimation and comparison of PESQ, ESTOI (%), and MOS_OVLR metrics among different enhancement models in the unseen test set 2.

System	Causal	PESQ	ESTOI (%)	MOS_OVLR
Noisy		1.96	49.97	1.87
LSTM	✓	2.65	70.64	2.64
CRN	✓	2.69	71.31	2.67
GCRN	✓	2.86	77.42	2.76
ConvTasNet	✓	2.73	76.63	2.74
CompNet	✓	**2.86**	**77.87**	**2.81**
BLSTM	×	2.85	76.28	2.89
BCRN	×	2.88	76.51	2.88
BGCRN	×	**3.07**	80.41	2.89
ConvTasNet_noncau	×	3.05	**82.56**	**3.11**
CompNet_noncau	×	3.03	81.54	2.94

In the seen test set 2, as shown in Table 5, both causal and non-causal systems could effectively remove background noise. Our proposed causal CompNet outperforms other causal baselines in all evaluation metrics. The non-causal CompNet performs worse than the non-causal ConvTasNet in the ESTOI and MOS_OVLR metrics but performs best in the PESQ metric.

In the evaluation set with unseen speakers, as shown in Table 6, we observed that all DNN-based speech enhancement models were able to effectively remove noise from untrained speakers under various conditions. Similar to the findings in Table 5, CompNet outperformed the other baseline models in all three evaluation metrics. However, in the context of non-causal systems, CompNet demonstrates the second highest performance in terms of ESTOI and MOS_OVLR.

Based on the analysis of the experimental results mentioned above, we conclude that the proposed CompNet effectively exploits the complementary characteristics to uncover the known information in causal

Table 7
Comparisons on the VoiceBank + Demand dataset.

Method	Year	WB-PESQ	CSIG	CBAK	COVL
Noisy		1.97	3.34	2.44	2.63
SEGAN (Pascual et al., 2017)	2017	2.16	3.48	2.94	2.80
MMSE-GAN (Chen et al., 2016)	2018	2.53	3.80	3.12	3.14
Wave-U-Net (Macartney & Weyde, 2018)	2018	2.40	3.52	3.24	2.96
MetricGAN (Fu et al., 2019)	2019	2.86	3.99	3.18	3.42
DCCRN (Hu, Liu, et al., 2020)	2020	2.68	3.88	3.18	3.27
TFT-Net (Tang et al., 2020)	2020	2.75	3.93	3.44	3.34
SE -Flow- μ (Strauss & Edler, 2021)	2021	2.43	3.77	3.12	3.09
SEAMNET (Borgström & Brandstein, 2021)	2021	–	3.87	3.16	3.23
DBT-Net (Yu et al., 2022)	2022	**3.30**	**4.59**	**3.75**	**3.92**
S-DCCRN (Lv et al., 2022)	2022	2.84	4.03	3.43	2.97
DCTCN (Zou & Zhu, 2021)	2022	2.83	3.91	3.37	3.37
CPTNN (Wang, He, & Zhu, 2022)	2022	3.07	4.40	3.59	3.76
CompNet	2023	2.90	4.16	3.37	3.53

systems. Therefore, this framework holds significant research significance in the context of causal systems. Furthermore, the complementary network also exhibits remarkable performance in non-causal systems.

5.4. Comparisons with baselines on VoiceBank + demand benchmark

In addition to the above experiments, we also carried out experiments on the VoiceBank + Demand dataset. The CompNet were compared with other baselines, see Table 7. Note that the causal CompNet was used in the training process.

From Table 7, it can be seen that CompNet achieves decent overall performance across these metrics, indicating the effectiveness of our approach. However, the performance of CompNet is not outstanding, which can be explained from two aspects. The first reason may be attributed to the extensive downsampling employed in our modeling, which may result in insufficient modeling of fine details (Yu et al., 2022). The second reason is that we did not incorporate highly complex modules at the specific module level, such as sub-band RNN (Chen, Wang, et al., 2022) or attention mechanisms (Wang, Cornell, et al., 2022). Exploring these module-level considerations is left as future work beyond the scope of this paper.

6. Discussion

The above experimental results indicate that our proposed CompNet, not only improves the performance of speech enhancement but also offers a novel perspective to explain the mechanisms underlying multi-domain networks. The following are some observed results.

Compared to time-domain networks and T–F domain methods, using a multi-domain approach yields better speech enhancement performance. While time-domain networks only enhance the speech waveform of the noisy speech, there may still exist residual noise in the distribution of the T–F domain, which differs from the target spectrum. Hence, we found that time-domain networks and time–frequency domain networks can benefit from each other, leading to the proposal of CompNet. At the same time, CompNet showcases a novel perspective for comprehending multi-domain networks.

Furthermore, through ablation experiments, it has been revealed that directly cascading the time-domain network and the T–F domain network is not the optimal approach. Therefore, we adhere to the "from coarse to fine" principle for speech enhancement. Specifically, we first utilize a time-domain network to estimate the speech waveform. Subsequently, we introduce a parallel branch in the T–F domain network to further refine the speech, enabling the attainment of optimal values for both magnitude and phase.

To summarize, our proposed CompNet for single-channel speech enhancement provides an explanatory framework for the mechanisms of multi-domain networks from a complementary perspective. Additionally, it follows the "from coarse to fine" principle, leading to better speech enhancement results.

7. Conclusion

This paper introduces a single-channel speech enhancement network called CompNet. CompNet initially utilizes a time-domain network to comprehensively enhance the waveform of noisy speech. Then, a dual-path network in the time–frequency domain is employed to further refine the estimated spectrum, aiming to approach the target speech more closely in the T–F domain. CompNet not only adheres to the "from coarse to fine" principle but also provides a different perspective to comprehend the improvements brought by multi-domain processing.

We conduct extensive experiments on the WSJ0-SI84 and VoiceBank + Demand datasets. The experimental results demonstrate the superiority of the network on both causal and non-causal systems. We plan to extend the proposed complementary estimation concept into other speech front-end tasks, e.g., acoustic echo cancellation, and multi-channel speech enhancement. Furthermore, our future research will take a more integrated and comprehensive approach, concentrating not only on further advancing complementary estimation methods but also on developing signal protection strategies. This is due to the negative impact of distortion on downstream tasks, such as ASR.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work is supported by the STI 2030—Major Projects (No. 2021ZD0201500), the National Natural Science Foundation of China (NSFC) (No. 61972437, No. 62201002), Distinguished Youth Foundation of Anhui Scientific Committee, China (No. 2208085J05), Special Fund for Key Program of Science and Technology of Anhui Province, China (No. 202203a07020008), Open Fund of Key Laboratory of Flight Techniques and Flight Safety, China, CACC (No. FZ2022KF15), the Open Research Projects of Zhejiang Lab, China (NO. 2021KH0AB06) and the Open Projects Program of National Laboratory of Pattern Recognition, China (NO. 202200014).

References

Abdelaziz, Ahmed Hussen, Gong, Yifan, & Stylianou, Yannis (2021). A fully convolutional recurrent network for real-time speech enhancement in the time domain. *IEEE Signal Processing Letters*, *28*, 119–123.

Bai, S., Kolter, J., & Koltun, V. (2018). Empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv:1803.01271.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *Transactions on Acoustics, Speech, and Signal Processing*, *27*(2), 113–120.

Borgström, B. J., & Brandstein, M. S. (2021). Speech Enhancement via Attention Masking Network (SEAMNET): An end-to-end system for joint suppression of noise and reverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 515–526.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing* (pp. 4960–4964).

Chen, J., Chen, J., Zhang, X., Hu, X., & Zhao, Y. (2022). Ideal ratio mask estimation with multi-domain collaboration for speech enhancement. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing* (pp. 3628–3632).

Chen, C., Feng, Y., Zhang, Z., Lu, W., & Liu, Y. (2022). A speech enhancement framework with hearing threshold prediction for hearing aids. *IEEE Transactions on Biomedical Engineering*, *69*(3), 812–819.

Chen, J., Wang, Y., oho, S. Y., Wang, D., & Healy, E. (2016). Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises. *Journal of the Acoustical Society of America*, *139*(5), 2604–2612.

Chen, J., Wang, Z., Tuo, D., Wu, Z., Kang, S., & Meng, H. (2022). FullSubNet+: Channel attention fullsubnet with complex spectrograms for speech enhancement. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing* (pp. 7857–7861).

Choi, H.-S., Park, S., Lee, J. H., Heo, H., Jeon, D., & Lee, K. (2021). Real-time denoising and dereverberation wtih tiny recurrent U-Net. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing* (pp. 5789–5793).

Chung, J., et al. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. Eprint Arxiv.

Deng, Xueqiang, Ye, Guoli, Song, Yan, et al. (2022). Self-attention-based deep neural network for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 3059–3071.

Fan, C., Tao, T., Liu, B., Yi, J., Wen, Z., & Liu, X. (2020). End-to-end post-filter for speech separation with deep attention fusion features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 1303–1314.

Fu, S.-W., Liao, C.-F., Tsao, Y., & Lin, S.-D. (2019). MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *Proc. ICML. PMLR* (pp. 2031–2041).

Guimarães, Heitor R., Nagano, Hitoshi, & Silva, Diego W. (2020). Monaural speech enhancement through deep wave-u-net. *Expert Systems with Applications*, *158*, Article 113582.

Guo, Y., & Li, C. (2023). An improved Wiener filtering method for speech enhancement based on spectral amplitude estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 235–245.

Han, K., Wang, Y., & Wang, D. L. (2014). Learning spectral mapping for speech dereverberation. In *Proc. int. conf. acoust. speech signal process.* (pp. 4628–4632).

Hochreiter, Sepp, & Schmidhuber, Jürgen (1997). Long short-term memory. *inproceedings*, *9*(8), 1735–1780.

Hu, Q., Fan, J., & Lu, X. (2022). A dual-path convolutional neural network for ideal ratio mask estimation in monaural speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, *30*, 297–312.

Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., et al. (2020). Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. In *Proc. Interspeech* (pp. 2472–2476).

Hu, Y., & Loizou, P. C. (2007). Evaluation of objective quality measures for speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *16*(1), 229–238.

Jensen, J., & Taal, C. (2016). An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(11), 2009–2022.

Kim, J. K., & Lee, H. G. (2023). Speech enhancement using ideal binary mask in noisy environments based on deep neural networks. *Applied Acoustics*, *184*(9930), 100.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Int. conf. learn. representations*.

Kleijn, W. B., et al. (2021). Generative speech coding with predictive variance regularization. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing* (pp. 6478–6482).

Koike-Akino, S., Babu, P., Zhang, L., Yoshioka, T., Sen, D., & Kawanishi, T. (2023). Hearing aid emulator: A speech separation approach. *IEEE Transactions on Biomedical Engineering*, *70*(1), 176–185.

Li, S., Li, D., Li, Y., & Li, Y. (2021). A convolutional recurrent neural network for ideal ratio mask estimation in speech enhancement. *IEEE Signal Processing Letters*, *28*, 155–159.

Li, A., Liu, W., Luo, X., et al. (2021). Deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network. In *ICASSP ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing* (ICASSP), (pp. 6628–6632). IEEE.

Li, A., Liu, W., Zheng, C., Fan, C., & Li, X. (2021). Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 1829–1843.

Li, A., Yuan, M., Zheng, C., & Li, X. (2020). Speech enhancement using progressive learning-based convolutional recurrent neural network. *Applied Acoustics*, *166*(7347), 100.

Li, A., Zheng, C., et al. (2021). On the importance of power compression and phase estimation in monaural speech dereverberation. *JASA Express Letters*, *1*(1), Article 014802.

Liu, Y., Chen, Z., Yang, J., & Xie, Y. (2022). Speech enhancement based on iterative spectral subtraction and non-stationary noise estimation. In *IEEE Signal Processing Letters 29*, 397–401.

Luo, Y., & Mesgarani, N. (2019). Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(8), 1256–1266.

Lv, Shubo, Fu, Yihui, Xing, Mengtao, Sun, Jiayao, Xie, Lei, Huang, Jun, et al. (2022). S-DCCRN: Super wide band DCCRn with learnable complex feature for speech enhancement. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing* (pp. 7767–7771). IEEE.

Macartney, C., & Weyde, T. (2018). Improved speech enhancement with the wave-U-Net. arXiv:1811.11307.

Naderi, B., & Cutler, R. (2021). Subjective evaluation of noise suppression algorithms in crowdsourcing. In *Proc. Interspeech* (pp. 2132–2136).

Olaf, Ronneberger, et al. (2015). U-Net: Convolutional networks for biomedical image segmentation. arXiv:1505.04597.

Paliwal, K., Wójcicki, K., & Shannon, B. (2011). The importance of phase in speech enhancement. *Speech Communication*, *53*(4), 465–494.

Pandey, A., & Wang, D. (2019). TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6875–6879).

Pascual, S., Bonafonte, A., & Serrà, J. (2017). Segan: Speech enhancement generative adversarial network. *2017*, In *Proc. Interspeech* (pp. 3642–3646).

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). Automatic differentiation in Pytorch.

Paul, D. B., & Baker, J. (1992). The design for the wall street journal-based CSR corpus. In *Proc. a workshop speech natural lang.* (pp. 23–26).

Qin, Xuebin, & Zhang, Zichen (2020). U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition*, *106*, Article 107404.

Recommendation ITU-T P ITU (2007). 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs. itu-telecommunication standardization sector.

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual Evaluation of Speech Quality (PESQ)-A new method for speech quality assessment of telephone networks and codecs. In *Proc. int. conf. acoust. speech signal process., vol. 2* (pp. 749–752).

Rouat, J. (2008). Auditory scene analysis: principles, algorithms, and applications (Wang, D. and Brown, G.J., Eds.; 2006) [Book review]. *IEEE Transactions on Neural Networks and Learning Systems*, *19*(1), 199.

Strauss, M., & Edler, B. (2021). A flow-based neural network for time domain speech enhancement. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing* (pp. 5754–5758).

Tan, K., & Wang, D. (2018). A convolutional recurrent neural network for real-time speech enhancement. In *Proc. Interspeech* (pp. 3229–3233).

Tan, K., & Wang, D. (2020). Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 380–390.

Tang, C., Luo, C., Zhao, Z., Xie, W., & Zeng, W. (2020). Joint time-frequency and time domain learning for speech enhancement. In *International joint conference on artificial intelligence*.

Valentini-Botinhao, Cassia, Wang, Xin, Takaki, Shinji, & Yamagishi, Junichi (2016). Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks. *Interspeech*.

Van Tasell, D. J (1993). Hearing loss, speech, and hearing aids. *Journal of Speech, Language, and Hearing Research*, *36*(2), 228–244.

Varga, A., & Steeneken, H. (1993). Assessment for automatic speech recognition: II. *Speech Communication*, *12*(3), 247–251.

Vincent, E., Sawada, H., Bofill, P., Makino, S., & Rosca, J. (2007). First stereo audio source separation evaluation campaign: data, algorithms and results. In *Proc. int. conf. ind. compon. anal. blind source separation* (pp. 552–559). Springer.

Wang, D., & Chen, J. (2018a). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1702–1726.

Wang, Z.-Q., Cornell, S., Choi, S., et al. (2022). TF-GridNet: Making time-frequency domain models great again for monaural speaker separation. arXiv:2209.03952.

Wang, K., He, B., & Zhu, W.-P. (2022). Cptnn: Cross-parallel transformer neural network for time-domain speech enhancement. In *2022 international workshop on acoustic signal enhancement* (IWAENC), Bamberg, Germany, (pp. 1–5).

Wang, Y., Narayanan, A., & Wang, D. (2014). On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(12), 1849–1858.

Wang, H., & Wang, D. (2022). Neural cascade architecture with triple-domain loss for speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 734–743.

Wang, Z.-Q., Wichern, G., & Le Roux, J. (2021). On the compensation between magnitude and phase in speech separation. *IEEE Signal Processing Letters*, *28*, 2018–2022.

Xie, Y., Xie, L., Chen, L., Zhang, J., & Li, B. (2022). A deep learning based speech enhancement method using ideal amplitude mask and noise adaptive loss. *IEEE Signal Processing Letters*, *29*, 1215–1219.

Xu, J., Xu, S., Kong, D., & Xu, Y. (2022). A multiscale and multitask U-Net model for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, *30*, 3885–3897.

Yin, D., Luo, C., Xiong, Z., et al. (2020). Phasen: A phase-and-harmonics-aware speech enhancement network. In *Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 05* (pp. 9458–9465).

Yu, G., Li, A., Wang, H., Wang, Y., Ke, Y., & Zheng, C. (2022). DBT-Net: Dual-branch federative magnitude and phase estimation with attention-in-attention transformer for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 2629–2644.

Zhang, Xinyi, Chen, Jingdong, Na, Xingyu, et al. (2022). Learning compact models for speech enhancement in the time domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 3376–3387.

Zhang, Q., Gao, X., & Liu, X. (2022). Complex-domain deep neural networks for speech enhancement based on a time-frequency domain representation. *Applied Acoustics*, *181*(8199), 100.

Zhang, X., Liu, H., Huang, Y., & Zhao, Y. (2023). Adversarial speech separation for multi-talker speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 63–74.

Zhang, W., Zhang, C., Zhao, D., & Du, J. (2023). A multi-bitrate multi-layer codec for speech coding in 5G communication networks. *IEEE Transactions on Audio, Speech, and Language Processing*, *31*, 1.

Zou, H., & Zhu, J. (2021). DCTCN: Deep complex temporal convolutional network for real time speech enhancement. In *2021 11th international conference on intelligent control and information processing* (pp. 112–118).