Applied Acoustics 194 (2022) 108792

Contents lists available at ScienceDirect

Applied Acoustics

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

SpecMNet: Spectrum mend network for monaural speech enhancement

Cunhang Fan [a], Hongmei Zhang [a], Jiangyan Yi [c,*], Zhao Lv [a,*], Jianhua Tao [c,d,*], Taihao Li [b], Guanxiong Pei [b], Xiaopei Wu [a], Sheng Li [e]

[a] Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China
[b] Artificial Intelligence Research Institute, Zhejiang Lab, Hangzhou 311121, China
[c] NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[d] University of Chinese Academy of Sciences, 100049 Beijing, China
[e] National Institute of Information and Communications Technology, Kyoto, Japan

ARTICLE INFO

Article history:
Received 24 November 2021
Received in revised form 14 March 2022
Accepted 22 April 2022
Available online 6 May 2022

Keywords:
Monaural speech enhancement
Speech distortion
Spectrum mend network
SI-SNR
BLSTM

ABSTRACT

Speech enhancement methods usually suffer from speech distortion problem, which leads to the enhanced speech losing so much significant speech information. This damages the speech quality and intelligibility. In order to address this issue, we propose a spectrum mend network (SpecMNet) for monaural speech enhancement. The proposed SpecMNet aims to retrieve the lost information by mending the weighted enhanced spectrum with weighted original spectrum. More specifically, the proposed algorithm consists of pre-enhancement network and the mend network. The main task of pre-enhancement network is to acquire the pre-enhanced spectrum so that it can remove the most of the noise signals. Because of the speech distortion problem, it loses a great deal of speech components. While the original spectrum has no speech information lost. Therefore, we utilize the original spectrum to mend the pre-enhanced spectrum by adding these two weighted spectrums so that the lost speech information can be retrieved. Then the mend network is used to predict mend weights for these two spectrums. Finally, the mended spectrum is used as the enhanced output. Our experiments are conducted on the TIMIT + (100 Nonspeech Sounds and NOISEX-92) datasets. Experimental results demonstrate that our proposed SpecMNet approach is effective to alleviate the speech distortion problem.

1. Introduction

In real scenarios, speech signals recorded by receivers are always contaminated by various noises, such as wind, machine, traffic and car noises. However, these noises can significantly degrade the performance of automatic speech recognition (ASR) [1], speech coding [2] and hearing aids [3]. To address this problem, speech enhancement is used to remove the noise from noisy speech signals. This paper focuses on the monaural speech denoising to improve the speech quality and intelligibility. In this paper, we only focus on the free field, which only includes the noise. The reverberation is not considered.

Over the past decades, many well-known traditional speech enhancement methods have been proposed, including Wiener filtering [4], spectral subtraction [5], statistical-based methods [6,7], and so on. Because of the powerful modeling ability of deep neural networks (DNNs), DNNs have been gradually applied to speech enhancement in recent years [8–13]. According to the output of network, speech enhancement methods can be divided into two categories in the time–frequency (T-F) domain: mapping based [14] and masking based [15]. The mapping based methods apply the spectral magnitude or log power spectrum as the network output to estimate the clean speech [14,5]. While the masking based methods use the mask as the network output, such as ideal binary mask (IBM) [16], ideal ratio mask (IRM) [17,15], ideal amplitude mask (IAM) [18] and phase sensitive mask (PSM) [19].

Because of the powerful modeling ability of recurrent neural networks (RNNs) and convolutional neural networks (CNNs), supervised speech enhancement acquires a good performance. To capture the long-term dependencies and make full use of the historical information of speech signals, in [19–23], authors utilize RNNs with long short-term memory (LSTM) or bidirectional LSTM (BLSTM) to speech enhancement. In addition, there are also many CNNs based speech enhancement methods [24–28]. Moreover, the time delay neural networks (TDNNs) [29] and convolutional recurrent networks (CRNs) [30–32] are applied to the speech

* Corresponding authors at: Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China (Zhao Lv); NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (Jianyang Yi and Jianhua Tao).

E-mail addresses: cunhang.fan@ahu.edu.cn (C. Fan), jiangyan.yi@nlpr.ia.ac.cn (J. Yi), kjlz@ahu.edu.cn (Z. Lv), jhtao@nlpr.ia.ac.cn (J. Tao), wxp2001@ahu.edu.cn (X. Wu), sheng.li@nict.go.jp (S. Li).

https://doi.org/10.1016/j.apacoust.2022.108792
0003-682X/

enhancement. Although above speech enhancement methods can improve the speech quality and intelligibility, they do not solve the speech distortion problem that is harmful for the performance of speech enhancement.

In order to describe the speech distortion problem, Fig. 1 illustrates the spectrogram example of a test speech sample. (a) is the spectrogram of noisy speech. (b) is the enhanced speech spectrogram by BLSTM, which is based on the IRM. (c) is the spectrogram of target clean speech. From Fig. 1 we can find that there are so many "leaks" in the enhanced speech spectrogram (marked in Fig. 1 (b) by bright boxes). These "leaks" are caused by the speech distortion. The reason is that mean squared error (MSE) loss function of the speech enhancement leads to generating over-smoothed spectrum that loses so many fine structures compared with the clean target speech [33]. However, these "leaks" lose a great deal of very important speech information so that damages the speech quality and intelligibility.

To deal with the speech distortion problem, in a preliminary study, we recently proposed a gated recurrent fusion (GRF) method to combine the original and enhanced features for robust end-to-end ASR [34]. The GRF aims to learn the raw fine structures from the original features and remove the noise signals from the enhanced features at the same time. Experimental results show that when the original and enhanced features are fused by the GRF, the ASR performance can be significantly improved. This proves that combining the original and enhanced speech can retrieve the lost speech information and relieve the speech distortion problem. However, the GRF utilizes another BLSTM layer to acquire deep representations, which increases the model parameters. Moreover, the main task of the GRF is to improve the performance of ASR not the speech quality and intelligibility.

In this study, we propose a **Spec**trum **M**end **Net**work (SpecMNet) to address the speech distortion problem. SpecMNet consists of pre-enhancement network and mend network. The pre-enhancement network is used to generate the pre-enhanced spectrum, which contains significant "leaks" so that loses much important speech information. While the original spectrum has no "leaks". Therefore, we apply the original spectrum to mend these "leaks" in the pre-enhanced spectrum. And the mend network is applied to predict the mend weight for each T-F bin. Then we utilize the mend weight to fuse the original and pre-enhanced spectrum. The fused spectrum is used as the finally enhanced output. Therefore, SpecMNet aims to mend "leaks" from the original and pre-enhanced spectrum so that it can retrieve the lost speech information. In addition, in order to overcome the shortcoming of the MSE loss, we fuse the MSE loss and scale-invariant source-to-noise ratio (SI-SNR) [35–37] loss to further improve the performance of speech enhancement.

The main contributions of this paper can be summarized as follows:

- To the best of our knowledge, this is the first work to apply the spectrum mend method to address the speech distortion problem.
- We propose the SpecMNet algorithm to mend the pre-enhanced spectrum by the weighted original spectrum so that it can retrieve the lost speech information.
- Experiments are conducted on TIMIT + (100 Nonspeech Sounds and NOISEX-92) datasets. Experimental results prove that our proposed SpecMNet method can significantly improve the performance of speech enhancement and it is effective to alleviate the speech distortion problem.

The rest of this paper is organized as follows. Section 2 presents the BLSTM based speech enhancement method. Section 3 introduces our proposed SpecMNet algorithm. We present the dataset and experimental setup in Section 4. Experimental results are shown in Section 5. The discussions are stated in Section 6. Section 7 draws conclusions.

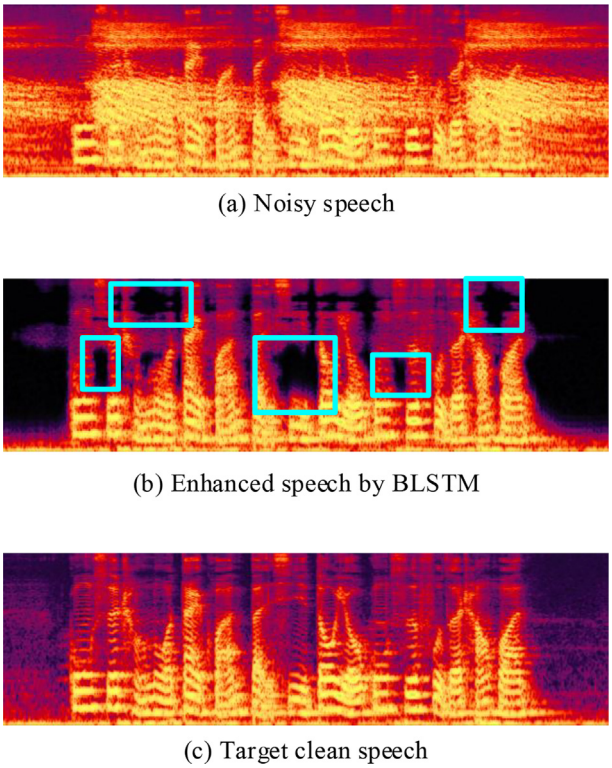

Fig. 1. The spectrogram example of a test speech sample. (a) The noisy speech. (b) The enhanced speech by BLSTM. (c) The target clean speech.

2. BLSTM based speech enhancement system

Monaural speech enhancement aims to remove the background noise n from the single channel noisy speech y. We model the noisy speech as follow:

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

where x denotes the target clean speech, t is the time sample index. The short-time Fourier transformation (STFT) can convert the speech signals in the time domain into the T-F domain, given as:

$$Y_{kf} = X_{kf} + N_{kf} \tag{2}$$

where Y, X and N are the corresponding STFT of y, x and n, respectively. k is the index of time frame and f is the index of frequency bin. For facilitating notations, the (k, f) is dropped in the rest of this paper.

The IAM is a widely-accepted mask for speech enhancement [18]. Therefore, we apply the IAM in this paper. The IAM is defined as follows:

$$M = \frac{|X|}{|Y|} \tag{3}$$

The noisy amplitude $|Y|$ is used as the input of BLSTM to estimate the IAM $\widehat{M}$.

$$\widehat{M} = \xi_{BLSTM}(|Y|) \tag{4}$$

where ξ_{BLSTM} denotes the mapping function of BLSTM. Then we can acquire the enhanced amplitude $|\hat{X}|$ by multiplying the estimated mask $\widehat{M}$ and the noisy input $|Y|$.

$$|\widehat{X}| = \widehat{M} \odot |Y| \tag{5}$$

where $\odot$ indicates element-wise multiplication.

Finally, the signal approximation (SA) [38] is used to compute the MSE between the enhanced and clean amplitude as the loss function:

$$J_{MSE} = \frac{1}{TF} \sum |||\hat{X}| - |X|||_F^2 \tag{6}$$

where T and F denote the number of time frames and frequency bins, respectively. The $||*||_F^2$ means the squared Frobenius norm.

3. Our proposed SpecMNet system

In order to retrieve the lost information caused by the speech distortion, this paper proposes a SpecMNet method for monaural speech enhancement. Fig. 2 illustrates the schematic diagram of our proposed SpecMNet method. From Fig. 2 we can find that the proposed SpecMNet method includes two sub-networks, namely pre-enhancement network and mend network. Firstly, the pre-enhancement network aims to acquire the pre-enhanced spectrum so that it can remove the most of the noise signals. However, this spectrum contains so many "leaks" leading to losing significant speech information. In order to address this problem, the mend network is proposed to estimate the mend weight. Then the mend weight is used to mend "leaks" and acquire the fused spectrum by adding the weighted pre-enhanced spectrum and the original spectrum. Therefore, the lost speech information can be retrieved by the fused spectrum. Finally, the fused spectrum is used as the enhanced output.

3.1. Pre-enhancement network

The main task of pre-enhancement network is to generate the pre-enhanced spectrum so that it can remove the most of noise signals. In a nutshell, we formulate the whole forward calculation process as follow:

$$h_{BLSTM} = \xi_{BLSTM}(|Y|), \tag{7}$$

$$\widehat{M} = ReLU(FNN(h_{BLSTM})), \tag{8}$$

$$|\widehat{X}_{pre}| = \widehat{M} \odot |Y|. \tag{9}$$

where the h_{BLSTM} denotes the hidden state of BLSTM, the *FNN* means the feed-forward neural network (FNN), the *ReLU* is the activation function of Rectified Liner Unit (ReLU) and $|\widehat{X}_{pre}|$ is the pre-enhanced spectrum.

The pre-enhancement network is same as the BLSTM based speech enhancement system described in Section II. Although pre-enhancement network can remove the most of noise signals, the enhanced spectrogram contains so many "leaks" and loses some very important speech components, which is caused by the speech distortion problem.

3.2. Mend network

In order to address speech distortion problem and retrieve the lost speech information, we propose the mend network to mend these "leaks" by fusing the weighted pre-enhanced spectrum and the weighted original spectrum. The red dotted box in Fig. 2 shows the schematic diagram of the mend network.

The main task of the mend network is to acquire the mend weight. We apply another BLSTM and FNN layer after the pre-enhancement network hidden layer, firstly. Then the sigmoid activation function is used to generate the mend weight, which limits the weight values ranging from 0 to 1. The process of mend network is as follow:

$$h_{mend} = FNN(\xi_{BLSTM}(h_{BLSTM})) \in \mathbb{R}^{T \times F}, \tag{10}$$

$$\begin{aligned} \lambda &= \sigma(h_{mend}) \in \mathbb{R}^{T \times F} \\ &= \frac{1}{1+e^{-h_{mend}}}. \end{aligned} \tag{11}$$

where T and F denote the number of time frame and frequency bin, respectively. h_{mend} is the hidden state of mend network. σ means the sigmoid activation function. λ is the mend weight.

We use λ as the mend weight of pre-enhanced spectrum. Although the original spectrum has no "leaks", it is contaminated by the noise signals. Therefore, we use $(1-\lambda)$ as the mend weight of original spectrum to mask noise signals in some degree. Finally, we mend these "leaks" by adding the weighted original spectrum and pre-enhanced spectrum.

$$|\widehat{X}| = \lambda \odot |\widehat{X}_{pre}| + (1-\lambda) \odot |Y| \tag{12}$$

Through the Eq. 12, these "leaks" in the pre-enhanced spectrum can be mended for each T-F bin so that the lost speech information by speech distortion can be retrieved. The mended spectrum $|\widehat{X}|$ is used as the output.

Finally, the inverse STFT (ISTFT) is used to convert the enhanced spectrum $|\widehat{X}|$ into time domain $\hat{x}$ (we drop the time sample index t):

$$\hat{x} = ISTFT\left(|\widehat{X}|; \Phi_y\right) \tag{13}$$

where Φ_y denotes the noisy phase spectrum.

3.3. Loss function

As for the pre-enhancement network, the SA is used to compute the MSE between the pre-enhanced and clean spectrum as the loss function:

$$J_{MSE}^{pre} = \frac{1}{TF} \sum |||\widehat{X}_{pre}| - |X|||_F^2 \tag{14}$$

Because the MSE loss leads to generating over-smoothed spectrum that loses so many fine structures, which is harmful for speech enhancement. In order to address this problem, instead of using the MSE loss, as for the mend network, we apply the SI-SNR as the loss function to improve the performance of speech enhancement.

The SI-SNR loss function is defined as follow:

$$x_{target} = \frac{\langle \hat{x}, x \rangle x}{\|x\|^2}, \tag{15}$$

$$e_{noise} = \hat{x} - x_{target}, \tag{16}$$

$$J_{SI-SNR} = 10log_{10} \frac{\|x_{target}\|^2}{\|e_{noise}\|^2}. \tag{17}$$

where $\|x\|^2 = \langle x, x \rangle$ is the signal power.

Finally, we fuse the MSE loss and SI-SNR loss with joint training framework as the total loss function of our proposed SpecMNet, which is defined as follow:

$$J = J_{MSE}^{pre} - \alpha J_{SI-SNR} \tag{18}$$

where α controls the weight of these two loss functions to balance the pre-enhancement network and the mend network. In this paper we set $\alpha = 0.1$.

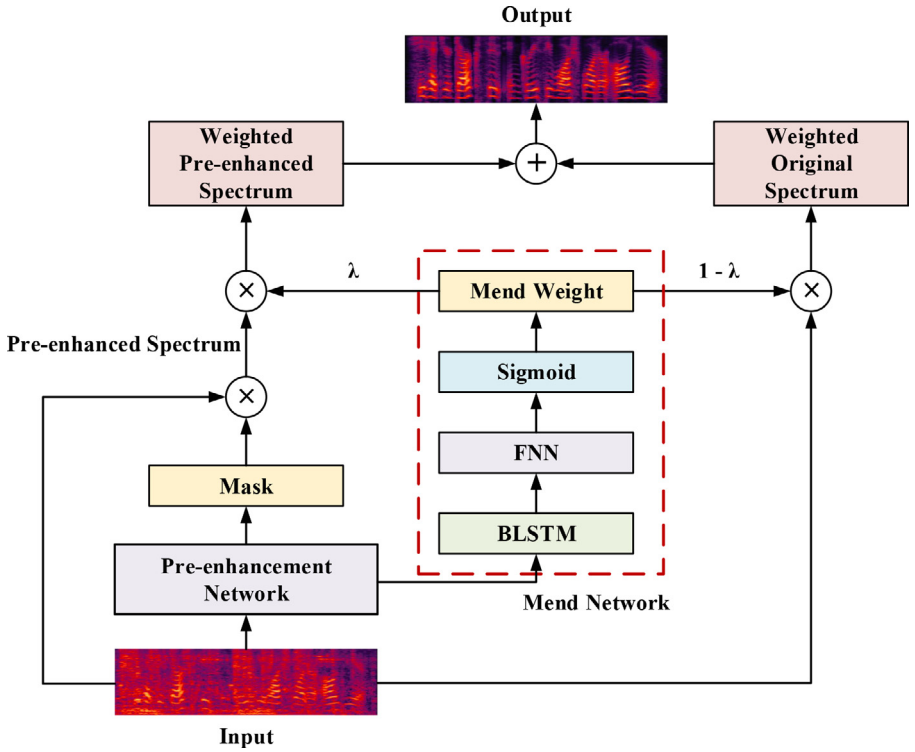

Fig. 2. The schematic diagram of our proposed SpecMNet method for monaural speech enhancement. The main task of the mend network is to predict the mend weight to fuse the original and pre-enhanced spectrum.

4. Dataset and experimental setup

4.1. Dataset

Our experiments are conducted on TIMIT database [39], which consists of 630 speakers each speaking 10 utterances. The training, development and test sets are created in the same manner. In order to acquire the noisy-clean pairs, the training and development sets use the 100 Nonspeech Sounds as the noise dataset, which includes 100 different noise types and can be download from the website.[1]

As for the training and development sets, they are generated by randomly selecting speakers and utterances from the TIMIT training set. Then they are mixed with the randomly selecting noise form the 100 Nonspeech Sounds. 6 different signal-to-noise ratios (SNRs) are used for training and development sets, which are −5, 0, 5, 10, 15 and 20 dB. As a result, the training and development sets have 21,726 and 6006 noisy-clean pairs, respectively. Detailed configuration is listed in Tables 1 and 2. As for the test set, besides the 100 Nonspeech Sounds, it also uses another twelve unseen noises, which are from the NOISEX-92 dataset [40]. In addition, the test set is generated by randomly selecting speakers and utterances from the TIMIT test set, which are mixed with noises at 6 different SNRs (−5, 0, 5, 10, 15 and 20 dB). As a result, the test set contains total 10,086 noisy-clean pairs. Detailed configuration is listed in Table 3.

Because the speakers and noises of development set are seen in the training set. We use the development set as the seen condition to evaluate different models. Instead, the speakers and NOISEX-92 dataset are unseen in the training set. We use the test set as the unseen condition.

[1] http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html.

Table 1
Configurations used for simulating training data.

Dataset	randomly selecting speakers and utterances from the TIMIT training set
Noise database	100 Nonspeech Sounds
SNR(dB)	−5, 0, 5, 10, 15 and 20
Number of utterance	21726

Table 2
Configurations used for simulating development data.

Dataset	randomly selecting speakers and utterances from the TIMIT training set
Noise database	100 Nonspeech Sounds
SNR(dB)	−5, 0, 5, 10, 15 and 20
Number of utterance	6006

Table 3
Configurations used for simulating test data.

Dataset	randomly selecting speakers and utterances from the TIMIT test set
Noise database	100 Nonspeech Sounds and NOISEX-92
SNR(dB)	−5, 0, 5, 10, 15 and 20
Number of utterance	10086

4.2. Our proposed SpecMNet model

The sampling rate of all generated waveform is 8000 Hz. We apply the 129 dimension noisy amplitude as the input feature. As for the STFT, the length of length hamming window is 32 ms and the window shift is 16 ms. The clean target amplitude is generated in the same manner.

As for the pre-enhancement network, there are two BLSTM layers and have 1024 units for every layer. As for the mask layer, we apply the ReLU as the activation function. And as for the mend network, there is only one BLSTM layer with 1024 units. And the sigmoid activation function is used to estimate the mend weight. Therefore, our proposed SpecMNet contains 3 BLSTM layers with 1024 units, totally.

4.3. Baseline models

For comparison purposes, two types of systems are built as two baselines in this paper.

- **BLSTM**: The BLSTM baseline means that we apply the pre-enhancement network for speech enhancement. Different from pre-enhancement network, the BLSTM baseline has 3 total BLSTM layers with 1024 units, which is same as our proposed SpecMNet method. We use the MSE J_{MSE} as the loss function of BLSTM baseline.
- **BLSTM-SISNR**: Compared with the BLSTM baseline, the only difference of BLSTM-SISNR is the loss function. The loss function of BLSTM-SISNR baseline contains SI-SNR loss:

$$J_1 = J_{MSE} - \alpha J_{SI-SNR} \tag{19}$$

In this paper, wee initialize the learning rate of all models are 0.0006. In addition, all models are optimized with the Adam algorithm [41].

4.4. Evaluation metrics

In this paper, to quantitatively evaluate speech enhancement results, the following objective measures are used: the perceptual evaluation of speech quality (PESQ) [42] measure, the extend short-time objective intelligibility (ESTOI) measure [43], the signal-to-distortion ratio (SDR) measure [35], overall Signal-to-Noise Ratio (ovlSNR) and Segmental Signal-to-Noise Ratio (segSNR) [44,45].

The PESQ aims to evaluate the speech quality, which ranges from −0.5 to 4.5. ESTOI extends from STOI [46] and ranges from 0 to 1, which aims to evaluate the speech intelligibility. SDR is the blind source separation (BSS) evaluation score. ovlSNR and segSNR are from 0 to ∞. For these evaluation metrics, the higher scores are, the better performances are.

5. Experimental results

Tables 4, 5, 6, 8 and 7 show the results of PESQ, ESTOI, SDR, segSNR and ovlSNR for different speech enhancement methods. "seen" case means that we evaluate different enhancement models on the development set. And "unseen" case means that we evaluate different enhancement models on the test set. "AVG." means the average value for all of the test SNR conditions.

5.1. Evaluation of MSE and SI-SNR loss

From Tables 5 and 6, we can find that when the MSE loss is replaced by the SI-SNR loss, the performance of speech enhancement can be improved for ESTOI and SDR evaluation metrics. More specifically, compared with BLSTM model, the BLSTM-SISNR model can acquire 1.15% increment in seen ESTOI, 1.05% increment in unseen ESTOI, 0.82 dB increment in seen SDR and 0.77 dB increment in unseen SDR for AVG. condition. These results show that the SI-SNR loss can improve the speech objective intelligibility. Because the SI-SNR and SDR are the BSS-eval scores. The BLSTM-SISNR model optimizes the SI-SNR directly so that it can improve the SDR performance. These results indicate that the SI-SNR loss is beneficial to the speech objective intelligibility and the SDR evaluation metric.

From Table 4 (PESQ evaluation metric), we can find that although the BLSTM-SISNR can acquire better performances than BLSTM for the AVG. and high SNRs, the BLSTM-SISNR gets worse results than BLSTM for the low SNRs (−5 and 0 dB). These results show that for high SNRs, the SI-SNR loss can improve the speech perceptual quality, but for the low SNRs it has worse performances than the MSE loss.

However, from Tables 8 and 7 (segSNR and ovlSNR evaluation metrics) we can find opposite results compared with Table 4. As for segSNR and ovlSNR evaluation metrics, compared with BLSTM, the BLSTM-SISNR has worse results for high SNRs and the AVG. condition. However, the BLSTM-SISNR can acquire better performances than BLSTM for low SNRs (−5 and 0 dB) no matter seen and unseen conditions. These results show that compared with MSE loss, the SI-SNR loss has a stronger capability to improve the SNR value of enhanced speech for the low SNRs noisy condition.

To summarize, from the above experimental results, we can know that SI-SNR loss and MSE loss have their own advantages and disadvantages. The SI-SNR loss only aims to optimize the BSS-eval scores so that leads to the worse performance for segSNR and ovlSNR evaluation metrics. While the MSE loss leads to generating over-smoothed spectrum that loses so many fine structures, which has poor performance for ESTOI and SDR evaluation metrics. However, whether BLSTM-SISNR or BLSTM model does not solve the speech distortion problem, which leads to the spectrum containing so many "leaks" and losing so much important speech information.

5.2. Comparison of our proposed SpecMNet model with baseline models

From Tables 4, 5, 6, 8 and 7, we can find that our proposed SpecMNet model acquires the best performance compared with baseline models (BLSTM and BLSTM-SISNR) for all SNRs no matter seen and unseen conditions. More specifically, compared with BLSTM baseline, our proposed SpecMNet model can get 0.15/0.15 (seen/unseen), 1.86%/1.66%, 1.20/1.12 dB, 1.03/0.94 dB and 0.80/0.73 dB improvements in PESQ, ESTOI, SDR, segSNR and ovlSNR evaluation metrics for AVG. condition. In addition, compared with BLSTM-SISNR baseline, our proposed SpecMNet model can acquire 0.09/0.09 (seen/unseen), 0.71%/0.61%, 0.38/0.35 dB, 1.37/1.35 dB and 2.01/1.97 dB improvements in PESQ, ESTOI, SDR, segSNR and ovlSNR evaluation metrics for AVG. condition. These results indicate the effectiveness of our proposed method.

Moreover, from these results we can make some valuable observations.

Firstly, compared with baseline models, our proposed SpecMNet model can obtain the best performance for all cases. The reason is that enhanced speech is always affected by the speech distortion problem, which leads to the spectrum containing so many "leaks" and losing so much very important speech information. However, speech distortion is harmful to the speech quality and intelligibility. In order to address the speech distortion problem, our proposed SpecMNet consists of pre-enhancement network and mend network. The main purpose of SpecMNet is to mend these "leaks" by fusing original and pre-enhanced spectrum so that it can retrieve the lost speech information. Therefore, our proposed SpecMNet model can acquire the best performance. Meanwhile, these results prove that our proposed method is effective for speech distortion problem.

Table 4
The objective results of PESQ for different speech enhancement methods. "seen" case means that we evaluate different enhancement models on the development set. And "unseen" case means that we evaluate different enhancement models on the test set. "AVG." means the average value for all of the test SNR conditions.

	Metric	PESQ						
	Test SNR	−5 dB	0 dB	5 dB	10 dB	15 dB	20 dB	AVG.
seen	Noisy	1.66	1.93	2.22	2.52	2.84	3.16	2.39
	BLSTM	2.55	2.78	2.97	3.14	3.30	3.47	3.03
	BLSTM-SISNR	2.45	2.76	3.02	3.25	3.45	3.62	3.09
	SpecMNet(ours)	**2.61**	**2.88**	**3.12**	**3.32**	**3.50**	**3.68**	**3.18**
unseen	Noisy	1.62	1.91	2.21	2.52	2.84	3.16	2.38
	BLSTM	2.50	2.74	2.94	3.12	3.29	3.46	3.01
	BLSTM-SISNR	2.40	2.72	3.00	3.23	3.44	3.61	3.07
	SpecMNet(ours)	**2.55**	**2.84**	**3.09**	**3.30**	**3.49**	**3.66**	**3.16**

Table 5
The objective results of ESTOI for different speech enhancement methods. "seen" case means that we evaluate different enhancement models on the development set. And "unseen" case means that we evaluate different enhancement models on the test set. "AVG." means the average value for all of the test SNR conditions.

	Metric	ESTOI(%)						
	Test SNR	−5 dB	0 dB	5 dB	10 dB	15 dB	20 dB	AVG.
seen	Noisy	44.13	55.97	67.90	78.73	87.42	93.44	71.26
	BLSTM	68.32	77.01	83.78	88.99	92.91	95.68	84.45
	BLSTM-SISNR	69.07	78.29	85.30	90.45	94.07	96.44	85.60
	SpecMNet(ours)	**69.98**	**79.26**	**86.15**	**91.13**	**94.54**	**96.77**	**86.31**
unseen	Noisy	43.34	55.11	67.07	78.03	86.90	93.10	70.59
	BLSTM	66.25	75.38	82.58	88.18	92.36	95.36	83.35
	BLSTM-SISNR	66.67	76.43	83.98	89.60	93.55	96.16	84.40
	SpecMNet(ours)	**67.51**	**77.26**	**84.70**	**90.18**	**93.98**	**96.44**	**85.01**

Table 6
The objective results of SDR for different speech enhancement methods. "seen" case means that we evaluate different enhancement models on the development set. And "unseen" case means that we evaluate different enhancement models on the test set. "AVG." means the average value for all of the test SNR conditions.

	Metric	SDR (dB)						
	Test SNR	−5 dB	0 dB	5 dB	10 dB	15 dB	20 dB	AVG.
seen	Noisy	−4.61	1.94	5.13	10.12	15.12	20.13	7.68
	BLSTM	9.67	12.19	14.81	17.58	20.52	23.66	16.41
	BLSTM-SISNR	10.54	13.04	15.62	18.38	21.33	24.46	17.23
	SpecMNet(ours)	**10.87**	**13.39**	**15.99**	**18.77**	**21.75**	**24.90**	**17.61**
unseen	Noisy	−4.61	1.96	5.14	10.12	15.12	20.13	7.68
	BLSTM	9.37	12.03	14.74	17.59	20.58	23.75	16.34
	BLSTM-SISNR	10.12	12.79	15.50	18.34	21.35	24.54	17.11
	SpecMNet(ours)	**10.41**	**13.11**	**15.84**	**18.71**	**21.75**	**24.95**	**17.46**

Table 7
The objective results of ovlSNR for different speech enhancement methods. "seen" case means that we evaluate different enhancement models on the development set. And "unseen" case means that we evaluate different enhancement models on the test set. "AVG." means the average value for all of the test SNR conditions.

	Metric	ovlSNR (dB)						
	Test SNR	−5 dB	0 dB	5 dB	10 dB	15 dB	20 dB	AVG.
	Noisy	−5.00	0.00	5.00	10.00	15.00	20.00	7.50
	BLSTM	8.96	11.23	13.58	15.90	18.03	19.79	14.58
	BLSTM-SISNR	9.31	11.33	13.17	14.67	15.65	16.13	13.37
seen	SpecMNet(ours)	**9.90**	**12.19**	**14.50**	**16.72**	**18.70**	**20.24**	**15.38**
	Noisy	−5.00	0.00	5.00	10.00	15.00	20.00	7.50
	BLSTM	8.71	11.10	13.51	15.88	18.04	19.82	14.51
	BLSTM-SISNR	8.97	11.12	13.07	14.64	15.67	16.18	13.27
unseen	SpecMNet(ours)	**9.53**	**11.95**	**14.35**	**16.66**	**18.68**	**20.25**	**15.24**

Secondly, from Table 4, we can find that as for PESQ evaluation metric, the BLSTM gets worse results for high SNRs while the BLSTM-SISNR has worse results for low SNRs. In addition, from Tables 8 and 7, we can find that, as for segSNR and ovlSNR evaluation metrics, the BLSTM gets worse performances for low SNRs while the BLSTM-SISNR has worse performances for high SNRs. While our proposed SpecMNet model can acquire superior performances whether low or high SNRs. This is because that SpecMNet applies the MSE loss at the pre-enhancement network, firstly. But the MSE loss leads to generating over-smoothed spectrum. In order to address this problem, SpecMNet utilizes the SI-SNR loss at the mend network. Finally, the joint training framework is used to combine these two losses deeply so that it can learn from these two losses' strong points and close the gap.

Table 8
The objective results of segSNR for different speech enhancement methods. "seen" case means that we evaluate different enhancement models on the development set. And "unseen" case means that we evaluate different enhancement models on the test set. "AVG." means the average value for all of the test SNR conditions.

	Metric	segSNR (dB)						
	Test SNR	−5 dB	0 dB	5 dB	10 dB	15 dB	20 dB	AVG.
seen	Noisy	−4.64	−2.01	1.11	4.60	8.37	12.29	3.29
	BLSTM	3.50	5.18	6.94	8.75	10.55	12.24	7.86
	BLSTM-SISNR	4.02	5.53	6.99	8.36	9.59	10.62	7.52
	SpecMNet(ours)	**4.43**	**6.22**	**8.02**	**9.83**	**11.60**	**13.25**	**8.89**
unseen	Noisy	−4.69	−2.05	1.08	4.59	8.37	12.31	3.27
	BLSTM	3.29	5.05	6.88	8.76	10.60	12.33	7.82
	BLSTM-SISNR	3.71	5.31	6.87	8.32	9.60	10.67	7.41
	SpecMNet(ours)	**4.11**	**5.98**	**7.86**	**9.75**	**11.58**	**13.26**	**8.76**

In order to show the experimental results more intuitively, we use three mean opinion score (MOS)-based evaluation metrics [47]: CSIG, CBAK and COVL, which are from 1 to 5. Where CSIG measures only the speech distortion. CBAK measures the interference of background noise. And COVL measures the overall effect. Fig. 3 shows the line chart of CSIG, CBAK and COVL for different speech enhancement methods under different SNRs. Where (a) shows the CSIG results, (b) shows the CBAK results and (c) shows the COVL results. From Fig. 3, we can know that although the BLSTM-SISNR acquires poor performances for low SNRs, our proposed SpecMNet can get comparable results to BLSTM baseline. In addition, as for the high SNRs, our proposed SpecMNet can acquire the best performances than baselines. These results suggest that fusing the original and pre-enhanced spectrum to mend "leaks" is effective for retrieving the lost speech information and speech distortion problem.

We chose 6 different types of noise, which consists of 3 seen noise (n3, n4 and n5 in 100 Nonspeech sounds) and 3 unseen noise (factory1, pink and white in NOISEX-92). Table 9 shows the objective results of PESQ for different noise types. From Table 9 we can find that the PESQ performances of seen noise are better than the unseen noise. This because that the enhanced model is trained by the seen noise. However, the unseen noise is unknown for the enhanced model. Our proposed SpecMNet can acquire the best performance no matter seen or unseen noise than baselines. These results indicate that our proposed method is effective for speech enhancement.

In addition, from Table 9 we can also find that speech enhancement methods have different performance for different noise types. For example, as for SpecMNet method, compared with the noisy input speech, the enhanced speech can acquire 0.75 improvement for the white noise, the pink noise is 0.6 and the factory1 is 0.43. These results suggest that the enhanced model can obtain a better performance for the white noise. This is because that the distinction between speech and white noise is relatively large. Therefore, the enhanced model can remove the noise very well.

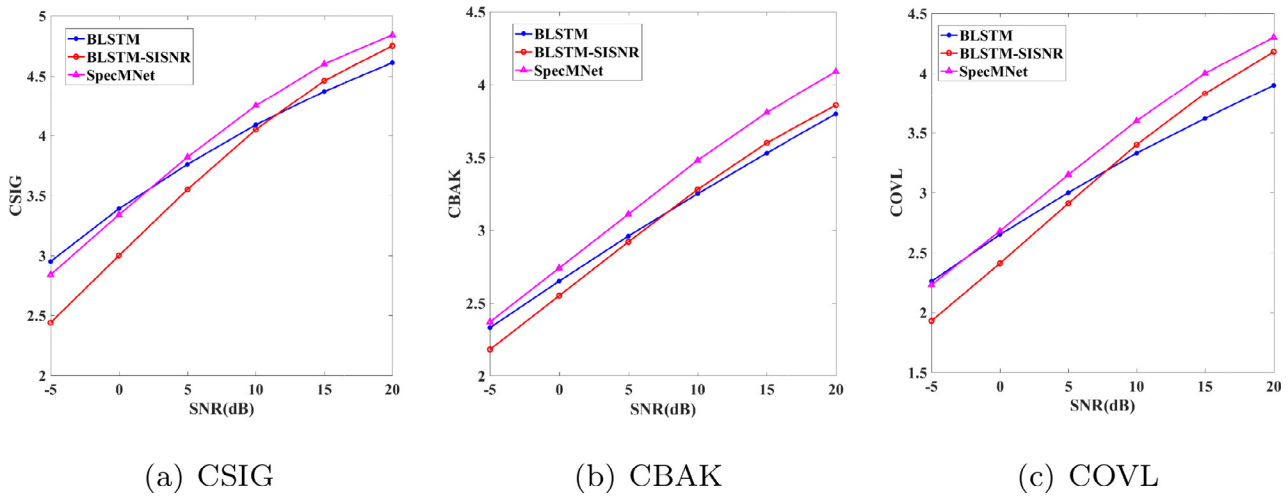

Fig. 3. The results of CSIG, CBAK and COVL for different speech enhancement methods under different SNRs. (a) CSIG results (b) CBAK results (c) COVL results.

Table 9
The objective results of PESQ for different noise types.

Methods	PESQ(AVG.)					
	seen			unseen		
	n3	n4	n5	factory1	pink	white
Noisy	2.03	2.21	2.18	2.31	2.22	1.94
BLSTM	2.74	2.80	2.96	2.65	2.68	2.60
BLSTM-SISNR	2.73	2.78	3.09	2.67	2.73	2.61
SpecMNet(ours)	**2.88**	**2.92**	**3.21**	**2.74**	**2.82**	**2.69**

6. Discussions

The above experimental results show that our proposed SpecMNet method by fusing the pre-enhanced and original spectrum to retrieve the lost speech information could alleviate the speech distortion quite well and acquire a pretty good performance. In addition, we can make some valuable observations as follows.

Our proposed SpecMNet method is effective to alleviate the speech distortion problem and can improve the performance of speech enhancement. The SpecMNet consists of pre-enhancement network and mend network. The pre-enhancement network aims to remove the most of noise signals, but it is affected by speech distortion, which has so many "leaks" and loses so much important speech information. In order to address this problem, the mend network is applied to fuse the original and pre-enhanced spectrum by the mend weight. Because the original spectrum has no "leaks" and no information lost but the noise signals. Therefore, we fuse the weighted original spectrum and weighted pre-enhanced spectrum to mend these "leaks" and retrieve the lost speech information so that the SpecMNet can alleviate the speech distortion problem.

The loss function of our proposed SpecMNet can learn from the MSE and SI-SNR losses' strong points and close the gap. From above experimental results, we can find that MSE and SI-SNR losses have their own advantages and disadvantages. For example, the MSE loss usually generates over-smoothed spectrum so that leads to the poor performance for ESTOI and SDR evaluation metrics. While the SI-SNR loss aims to optimize the BSS-eval scores, which has poor performance for segSNR and ovlSNR evaluation metrics. However, our proposed SpecMNet method can get superior performances for these evaluation metrics. This is because that we use the MSE loss for the pre-enhancement network and utilize the SI-SNR loss for mend network. Then the joint training framework is applied to combine these two losses. Therefore, the proposed SpecMNet can fuse these two losses deeply.

In summary, our proposed SpecMNet method could alleviate the speech distortion quite well. In addition, the loss function of SpecMNet is beneficial to improve the performance of speech enhancement.

7. Conclusion

In this paper, in order to address the speech distortion problem, we propose a spectrum mend network (SpecMNet) for monaural speech enhancement. Our proposed SpecMNet method includes two sub-networks, namely pre-enhancement network and mend network. The pre-enhancement network aims to acquire the pre-enhanced spectrum and remove the most of noise signals. And the main task of mend network is to predict mend weights for pre-enhanced and original spectrums. Because of the speech distortion, the pre-enhanced spectrum has so many "leaks" leads to losing so much speech information. Therefore, we apply the original spectrum to mend the pre-enhanced spectrum by adding these two weighted spectrums. Besides, we fuse the MSE loss and SI-SNR loss by joint training framework. Experiments on TIMIT + (100 Nonspeech Sounds and NOISEX-92) datasets demonstrate that our proposed SpecMNet method is effective to alleviate the speech distortion problem. In the future, we plan to extend our proposed approach to the complex spectral speech enhancement in order to avoid phase mismatch problem.

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.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2021ZD0201502), the National Natural Science Foundation of China (NSFC) (No.61972437), the Open Research Projects of Zhejiang Lab (NO. 2021KH0AB06) and the Open Projects Program of National Laboratory of Pattern Recognition (NO. 202200014).

References

[1] Li J, Deng L, Gong Y, Haeb-Umbach R. An overview of noise-robust automatic speech recognition. IEEE/ACM Trans Audio Speech Language Process 2014;22(4):745–77.
[2] Li J, Sakamoto S, Hongo S, Akagi M, Suzuki Y. Two-stage binaural speech enhancement with wiener filter for high-quality speech communication. Speech Commun 2011;53(5):677–89.
[3] Levitt H. Noise reduction in hearing aids: A review. J Rehab Res Develop 2001;38(1):111–22.
[4] Scalart P, et al. Speech enhancement based on a priori signal to noise estimation. In Proc. ICASSP, vol. 2. IEEE; 1996. pp. 629–632..
[5] Boll S. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans Acoust Speech Signal Process 1979;27(2):113–20.
[6] Ephraim Y, Malah D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process 1984;32(6):1109–21.
[7] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process 1985;33(2):443–5.
[8] Pandey A, Wang D. Dense cnn with self-attention for time-domain speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 2021;29:1270–9.
[9] Tan K, Wang D. Towards model compression for deep learning based speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 2021;29:1785–94.
[10] Pandey A, Wang D. On cross-corpus generalization of deep learning based speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 2020;28:2489–99.
[11] Li A, Liu W, Zheng C, Fan C, Li X. Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 2021.
[12] Li A, Yuan M, Zheng C, Li X. Speech enhancement using progressive learning-based convolutional recurrent neural network. Appl Acoust 2020;166:107347.
[13] Li A, Zheng C, Zhang L, Li X. Glance and gaze: A collaborative learning framework for single-channel speech enhancement. Appl Acoust 2022;187:108499.
[14] Xu Y, Du J, Dai L-R, Lee C-H. A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans Audio Speech Lang Process 2014;23(1):7–19.
[15] Wang Y, Narayanan A, Wang D. On training targets for supervised speech separation. IEEE/ACM Trans Audio Speech Lang Process 2014;22(12):1849–58.
[16] Roman N, Wang D, Brown GJ. Speech segregation based on sound localization. J Acoust Soc Am 2003;114(4):2236–52.
[17] Hummersone C, Stokes T, Brookes T. On the ideal ratio mask as the goal of computational auditory scene analysis. Blind source separation. Springer 2014:349–68.
[18] Wang D, Chen J. Supervised speech separation based on deep learning: An overview. IEEE/ACM Trans Audio Speech Lang Process 2018;26(10):1702–26.
[19] Erdogan H, Hershey JR, Watanabe S, Le Roux J. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. p. 708–12.
[20] Weninger F, Hershey JR, Le Roux J, Schuller B. Discriminatively trained recurrent neural networks for single-channel speech separation. In: 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP). IEEE. p. 577–81.
[21] Weninger F, Eyben F, Schuller B. Single-channel speech separation with memory-enhanced recurrent neural networks. 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE 2014:3709–13.
[22] Weninger F, Erdogan H, Watanabe S, Vincent E, Le Roux J, Hershey JR, Schuller B. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In: International conference on latent variable analysis and signal separation. Springer; 2015. p. 91–9.
[23] Fan C, Tao J, Liu B, Yi J, Wen Z. Joint training for simultaneous speech denoising and dereverberation with deep embedding representations. Proc. Interspeech 2020 2020:4536–40.
[24] Park SR, Lee JW. A fully convolutional neural network for speech enhancement. Proc. Interspeech 2017 2017:1993–7.
[25] Pascual S, Bonafonte A, Serrà J. Segan: Speech enhancement generative adversarial network. Proc. Interspeech 2017 2017:3642–6.
[26] Fan C, Liu B, Tao J, Yi J, Wen Z, Bai Y. Noise prior knowledge learning for speech enhancement via gated convolutional generative adversarial network. 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE 2019:662–6.

[27] Fu S-W, Tsao Y, Lu X. Snr-aware convolutional neural network modeling for speech enhancement. Interspeech 2016:3768–72.
[28] Tan K, Chen J, Wang D. Gated residual networks with dilated convolutions for monaural speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 2018;27(1):189–98.
[29] Fan C, Liu B, Tao J, Yi J, Wen Z, Song L. Deep time delay neural network for speech enhancement with full data learning. 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE 2021: 1–5.
[30] Naithani G, Barker T, Parascandolo G, Bramsl L, Pontoppidan NH, Virtanen T, et al. Low latency sound source separation using convolutional recurrent neural networks. 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE 2017:71–5.
[31] Takahashi N, Goswami N, Mitsufuji Y. Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation. 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE 2018:106–10.
[32] Tan K, Wang D. A convolutional recurrent neural network for real-time speech enhancement. Interspeech 2018:3229–33.
[33] Liu B, Nie S, Liang S, Liu W, Yu M, Chen L, Peng S, Li C. Jointly adversarial enhancement training for robust end-to-end speech recognition. INTERSPEECH 2019:491–5.
[34] Fan C, Yi J, Tao J, Tian Z, Liu B, Wen Z. Gated recurrent fusion with joint training framework for robust end-to-end speech recognition. IEEE/ACM Trans Audio Speech Lang Process 2020;29:198–209.
[35] Vincent E, Gribonval R, Févotte C. Performance measurement in blind audio source separation. IEEE Trans Audio Speech Lang Process 2006;14(4):1462–9.
[36] Luo Y, Mesgarani N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM Trans Audio Speech Lang Process 2019;27(8):1256–66.
[37] Fan C, Tao J, Liu B, Yi J, Wen Z, Liu X. End-to-end post-filter for speech separation with deep attention fusion features. IEEE/ACM Trans Audio Speech Lang Process 2020;28:1303–14.
[38] Huang P-S, Kim M, Hasegawa-Johnson M, Smaragdis P. Deep learning for monaural speech separation. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. p. 1562–6.
[39] Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett, Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1–1.1, NASA STI/Recon technical report n, vol. 93; 1993..
[40] Varga A, Steeneken HJ. Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun 1993;12(3):247–51.
[41] Kingma DP, Ba J. Adam: A method for stochastic optimization. Computer Science; 2014.
[42] Rix AW, Hollier MP, Hekstra AP, Beerends JG. Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part i–time-delay compensation. J Audio Eng Soc 2002;50(10):755–64.
[43] Jensen J, Taal CH. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. IEEE/ACM Trans Audio Speech Lang Process 2016;24(11):2009–22.
[44] Papamichalis PE. Practical Approaches to Speech Coding. USA: Prentice-Hall Inc; 1987.
[45] Hansen JH, Pellom BL. An effective quality evaluation protocol for speech enhancement algorithms. In: Fifth international conference on spoken language processing.
[46] Taal CH, Hendriks RC, Heusdens R, Jensen J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE. p. 4214–7.
[47] Hu Y, Loizou PC. Evaluation of objective quality measures for speech enhancement. IEEE Trans Audio Speech Lang Process 2007;16(1):229–38.