Speech Dereverberation Based on Scale-aware Mean Square Error Loss

Luya Qiang[1], Hao Shi[2⋆], Meng Ge[1(✉)], Haoran Yin[1], Nan Li[1], Longbiao Wang[1(✉)], Sheng Li[3], and Jianwu Dang[1,4]

[1] [1]Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China
{qiangluya, gemeng, longbiao_wang}@tju.edu.cn
[2] Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto, Japan
[3] National Institute of Information and Communications Technology (NICT), Kyoto, Japan
[4] Japan Advanced Institute of Science and Technology, Ishikawa, Japan

Abstract. Recently, deep learning-based speech dereverberation approaches have achieved remarkable performance by directly mapping the input spectrogram to a target spectrogram or time-frequency mask. However, these approaches are usually optimized under distance-related objective functions—the mean square error (MSE). The traditional MSE training criterion results in a strong inherent uniform variance statistical assumption on the target speech and noise during training, which cannot be satisfied in real-world scenarios. To alleviate such an assumption mismatch problem, we propose a speech dereverberation solution called Scale-aware Speech Dereverberation (SaSD) based on scaled-MSE. Specifically, we modify the MSE with different scales for each frequency band and progressively reduce the gap between the low- and high-frequency ranges to make the error follow the assumption of MSE assumption. Experiments demonstrated that SaSD achieved 1.0 SRMR and 0.8 PESQ improvements over the mapping baseline system.

Keywords: Speech dereverberation · Scale-aware mean square error · Progressive learning · Deep learning.

1 Introduction

In real-world environments, the sound reaching the ears comprises the clean direct-path speech and its reflections from various surfaces, which drastically reduce speech signal intelligibility [11]. To minimize the distortions in real-world cases, speech dereverberation, as an important front-end signal processing module, is designed to remove the adverse effects of reverberation for the back-end speech applications, such as speech recognition [10, 13].

Recently, some researchers have explored the use of deep neural networks for speech dereverberation, such as using spectral mapping methods [4, 7]. The key

⋆ Hao Shi is the joint first author. *Corresponding author.

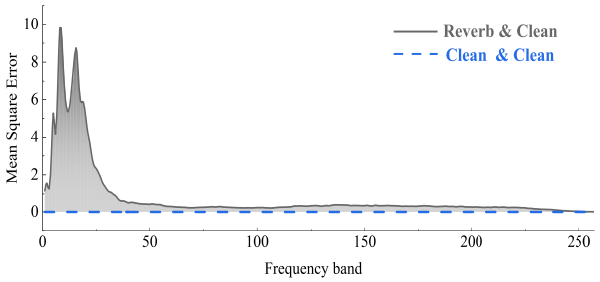

Fig. 1. The mean square error of the observed noisy speech and clean speech for all frequency bands on REVERB dataset.

strategy is to view the speech dereverberation as a regression problem, where the nonlinear regression function can be parametrized using deep neural networks. In recent years, there have been many models based on spectral mapping, such as DNN [24], CNN [14], RNN [23], GAN [15, 22], CRNN [20], etc.

Generally, these spectral mapping approaches are optimized by a distance-related objective function, such as the mean square error (MSE) [12, 3]. However, simple applying a distance-related objective function to dereverberation network training results in strong inherent assumptions on the statistics of the clean speech and noise [2, 5]. For example, the MSE objective function assumes that the errors of all frequency bands have zero means and uniform variance [19]. Unfortunately, this assumption cannot be met in real-world scenarios because the target clean speech and noisy speech have a non-uniform spectral distribution as shown in Fig. 1. From it, we can observe that the error between the target clean speech and noisy speech in the approximately 1–50 frequency bands are much larger than that in higher frequency range. Such an assumption mismatch for the MSE has the problem of underestimating the error in the frequency range with lower power, leads to the training difficulty of speech dereverberation in the higher frequency range.

To address the above problem, we propose a speech dereverberation approach based on scaled-MSE loss function called Scale-aware Speech Dereverberation (SaSD). To make the error follow the statistical assumption of the MSE, we first modify the MSE loss function using different weights for frequency bands, where the low-frequency bands are given larger weights and the high-frequency bands are given smaller weights. Additionally, motivated by the idea of SNR-based progressive learning (PL) in the literature [6, 21], we use the PL strategy to gradually reduce the gap between high- and low-frequency range. Thus, we can apply the MSE loss function with the help of the progressive learning approach to equally treat the error for each frequency. From the comparison of the loss curves, it can be observed that by using our method, the gap between high-frequency and low-frequency gradually decreases with the progress of the stage and significantly reduce the error between the predicted spectrogram and the target spectrogram. And from the experimental results, it is helpful to improve the speech dereverberation ability by using the method we proposed.

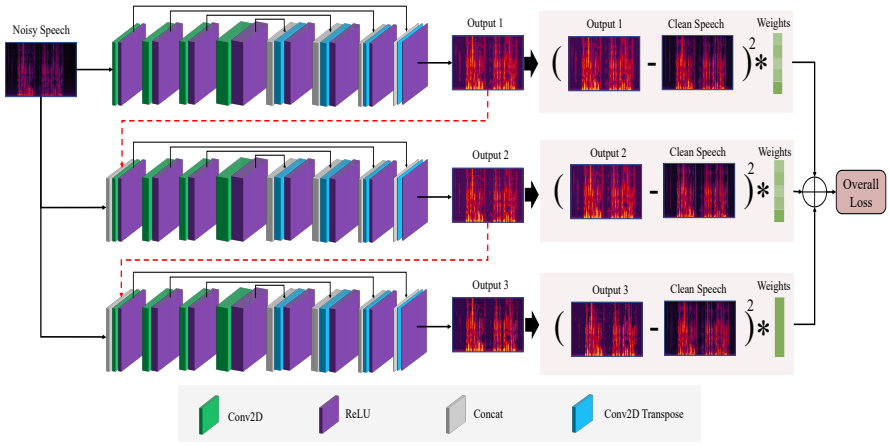

Fig. 2. The diagram of the SaSD. The whole system consists of three progressive stages. Each stage contains three FCN blocks and each block is a U-net structure, including four 2D convolution layers transposed convolution layers.

2 Spectral Mapping with MSE Loss

In real-world environments, the original source (direct sound) $s(t)$ is easily destroyed by convolutional noise $r(t)$ and additive background noise $b(t)$. Thus, the observed signal $y(t)$ can be written as follows:

$$y(t) = s(t) * r(t) + b(t) \tag{1}$$

Mapping-based methods aim to learn a nonlinear function $\mathcal{F}$ from the observed noisy speech $y(t)$ into the desired clean speech $s(t)$, as described by the following:

$$y(t) \xrightarrow{\mathcal{F}} s(t) \tag{2}$$

To learn $\mathcal{F}$, the neural network is trained to reconstruct the target speech spectrum $S(n, f)$ from the corresponding input noisy speech spectrum $Y(n, f)$ [1]. In traditional methods, the parameters of the model are determined by minimizing the mean square error (MSE) objective function as follows:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N * F} \sum_{n=1}^{N} \sum_{f=1}^{F} ||\mathcal{F}(Y(n, f)) - S(n, f)||^2 \tag{3}$$

where N and F are the number of frames and frequency bands, respectively. The n and f are the corresponding index of the frame and frequency band.

3 Scale-aware Speech Dereverberation Architecture

SaSD is a progressive mapping pipeline with multiple stages. We use a three-stage architecture in this work, as shown in Fig. 2. Different from the traditional MSE objective function, we use scaled-MSE for different frequency bands as the training criterion, and design three-stage progressive architecture to alleviate the assumptions in MSE loss.

3.1 Scaled-MSE loss

Since the MSE objective function assumes that the mean error of all frequencies are zero and the variance are the same, this assumption cannot be met in the fact that the same distortion in different frequency bands has different effects on speech quality. Motivated by this, we firstly apply a scaled-MSE loss to reduce the error gap between low- and high-frequency bands, which is defined as:

$$\mathcal{L}_{\text{Scaled-MSE}}(w_f) = \frac{1}{N * F} \sum_{n=1}^{N} \sum_{f=1}^{F} w_f ||\mathcal{F}(Y(n,f)) - S(n,f)||^2 \tag{4}$$

where w_f is the scale parameter of the f-th frequency band. Here, we use a linear function to define w_f, so the weight value will change as the frequency band changes. The formula is defined as follows:

$$w_f = 1 - \frac{1 - w_{min}}{F} * (f - 1) \tag{5}$$

where w_{min} represents the minimum scale threshold hyperparameter. It is used to control the minimum weight of the high frequency band in the training process.

By applying scaled-MSE training criterion, the low-frequency bands are given large weights and the high-frequency bands are weighted with smaller scales. During the training stage, the dereverberaton network pays more attention to reducing the reconstruction error of low-frequency bands, rather than treating each frequency band equally.

3.2 Progressive scaled-MSE loss

Reducing the gap between high and low frequency bands is a complex learning process, and direct mapping optimization with scaled-MSE training criterion is hard to achieve the expected goal. Motivated by the progressive learning study in speech enhancement [6], we propose the progressive scaled-MSE loss to decompose the complicated non-linear mapping problem into a series of easier sub-problems. The key idea is to gradually reduce the reconstruction error between noisy and clean speech at low frequencies in a multi-stage manner, and finally apply the original MSE objective function to guide the dereverberation network. The progressive scaled-MSE loss is define as follow:

$$\begin{aligned} \mathcal{L}_{\text{Prog-MSE}} &= \sum_{p=1}^{P} \alpha_p \mathcal{L}_{\text{Scaled-MSE}}(w_f^p) \\ &= \frac{1}{NF} \sum_{p=1}^{P} \sum_{n=1}^{N} \sum_{f=1}^{F} \alpha_p w_f^p ||\mathcal{F}(Y(n,f)) - S(n,f)||^2 \end{aligned} \tag{6}$$

where P denotes the number of stages in the whole system, and α_p represents the weight coefficient of the p-th stage loss. The w_f^p is calculated from the parameter w_{min} at p-th stage using Eq. (5). In this study, we apply three-stage mapping architecture as shown in Fig. 2.

4 Experiments and Discussion

4.1 Dataset

The experiments were conducted on the REVERB challenge dataset [10, 9]. The database contains simulated and real recordings, that have been sampled from different rooms with different reverberation levels and 20 dB SNR of background noise. The simulated data were generated by convolving the room impulse responses (RIRs) collected from rooms of three different sizes (small, medium, large) and two different microphone positions (near, far) by using single-channel microphones and clean speech utterances from WSJCAM0 [17]. The corpus was divided into training, validation and test sets. The training data included 7,861 simulated recordings, whereas the test data contained simulated and real recordings. The validation set used only the simulated data. All of the speech signals were sampled at 16 kHz.

4.2 Experimental setup

For all models, the window length and hop size were 32 ms and 16 ms, and the FFT length was 512. All of the models were implemented on TensorFlow, and the weights of them were randomly initialized. The architecture of SaSD was divided into three stages with FCN blocks [18]; each block is a U-net structure [25] that mainly consists of four two-dimensional (2D) convolution layers and four 2D transposed convolution layers, where the numbers of filters for each convolution layer were 8, 16, 16, 32, 16, 16, 8 and 1. ReLU was used as the activation function, as shown in Fig. 2. The α_p of our proposed model for stages 1 and 2 is 0.1, and the α_p for stage 3 is 1. In the experiments, the perceptual evaluation of speech quality (PESQ) [16] and speech-to-reverberation modulation energy ratio (SRMR) [8] were used as the evaluation metrics. All approaches use noisy phase information to reconstruct the enhanced waveform. To choose the model for speech reverberation, several models were compared on the REVERB dataset, which are described as follows:

Traditional methods: ①**Reverb**: reverberant spectrogram; ②**Mapping**: mapping system with one FCN block; ③**Naive Iteration**: mapping system with three naive iterative FCN blocks and use the final output to calculate the loss.

Proposed methods (SaSD): the proposed system that consists of three progressive stages. During the training, the intermediate predicted spectrogram from previous stage is concatenated with the noisy spectrogram as input into the next stage. The overall loss is the sum of the loss value in each stage. According to the change of the weight value, it is divided into two cases as:
①***hard***: the proposed approach that divides the entire frequency domain into three equal parts, with each part adopting a fixed weight at each stage; in this experiment, (1, 0.5, 0) and (1, 0.75, 0.5) were used in stages 1 and 2, respectively.
②***linear*** **(w_f^1- w_f^2- w_f^3)**: the approach in which the threshold of w_f for stages 1, 2, and 3 are w_f^1, w_f^2, and w_f^3, respectively. At each stage, as the frequency increases, the weight values of different frequencies linearly decrease from 1 to the threshold. In this experiment, we set a total of three sets of different weight thresholds to evaluate the performance of our proposed model.

Table 1. PESQ and SRMR in a comparative study on the REVERB dataset. In the table, w_{min} denotes the weight threshold of the high frequency band; *Mode* denotes the change in weight value; *hard* means each part adopting a fixed weight at each stage and *linear* means the weight values will linearly decrease from 1 to the threshold; *#1*, *#2*, *#3* denote the number of stages.

Model	Configurations				PESQ			SRMR					
	w_{min} (#stage)				Simulated			Simulated			Real		
	Mode	#1	#2	#3	Far	Near	Ave.	Far	Near	Ave.	Far	Near	Ave.
Reverb	-	-	-	-	2.15	2.59	2.37	3.43	3.94	3.68	3.19	3.17	3.18
Mapping	-	-	-	-	2.42	2.66	2.54	4.20	4.66	4.43	4.01	3.66	3.83
Naive Iteration	-	-	-	-	2.42	2.74	2.58	4.25	4.70	4.47	3.86	3.55	3.71
SaSD	hard	0	0.5	1	**2.46**	2.76	2.61	4.67	5.17	4.92	4.59	4.08	4.33
	linear	0	0.5	1	2.43	2.70	2.57	**4.72**	5.14	**4.93**	**5.03**	**4.62**	**4.82**
	linear	0.5	0.75	1	2.45	2.74	2.60	4.63	**5.19**	4.91	4.69	4.22	4.46
	linear	1	1	1	**2.46**	**2.78**	**2.62**	4.32	4.89	4.60	4.15	3.70	3.93

4.3 Experimental results and discussion

We compared our SaSD with previous baseline systems on REVERB in terms of PESQ and SRMR in Table 2. Regardless of the evaluation index used, the experimental results showed that the highest scores were obtained by SaSD. Additionally, when the average frequency was divided into three parts with equal weights, a high PESQ measure was obtained. Furthermore, on simulated data, the naive iteration mapping model achieved better performance than the single mapping model in terms of both PESQ and SRMR. Conversely, the opposite was true for real data. This outcome may be due to the difference between the data distributions of the real and simulated datasets, and such a superimposed network may be more affected than a single network, thereby resulting in performance degradation on the real dataset. However, when the direct mapping of noisy speech to pure speech was decomposed into multiple stages as the frequency weight threshold increased, the model also achieved good performance. To verify the effectiveness of the progressive learning strategy, the mean square error for all frequency bands at each stage is shown in Fig. 3. It is observed from Fig. 3 that the gap between the high-frequency and low-frequency results gradually decreases with the progress of the stage. This fact shows that progressive learning with scaled loss can help the dereverberation network alleviate the assumption mismatch of MSE. In addition, we further verified the effectiveness with naive iteration scheme, as shown in Fig. 4. Compared with the results under the naive iteration scheme, we found that our ProgressSD approach can significantly reduce the deviation of the predicted spectrogram from the target spectrogram. This improvement is from the greater stability and better convergence of the dereverberation network trained under scaled-MSE.

5 Conclusion and future work

In this paper, we proposed a system featuring scaled-MSE and PL, which is called SaSD. We modified the MSE with different scales for each frequency band

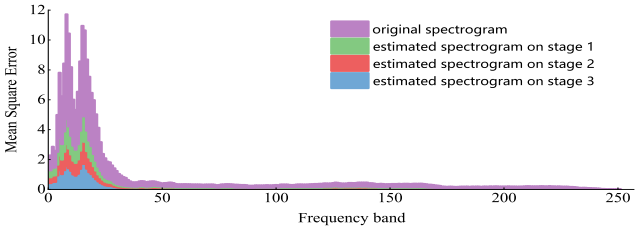

Fig. 3. Scaled mean square error loss values at each frequency range for different progressive stages.

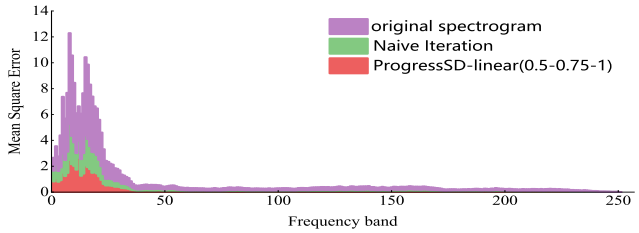

Fig. 4. Mean square error loss values at each frequency band for our SaASD and 3-mapping naive iteration system.

and progressively reduced the gap between the low- and high-frequency range that was used for speech dereverberation based on mapping; this was done in order to solve the problem of the nonuniform variance of different frequency bands that makes some regions in the spectrogram difficult to learn. The experimental results showed that the loss curve became more stable and showed better convergence with SaSD. It was also found that all the approaches of PL that used scaled-MSE exhibited improved performance, particularly with respect to SRMR and PESQ. In the future, we will replace our current network with a state-of-the-art network structure.

References

1. Avargel, Y., Cohen, I.: System identification in the short-time fourier transform domain with crossband filtering. IEEE TASLP **15**(4) (2007)
2. Cho, E., Lee, B., Schafer, R., Widrow, B.: Single channel speech enhancement using outlier detection. arXiv preprint arXiv:1605.01329 (2016)
3. Erdogan, H., Hershey, J.R., Watanabe, S., Le Roux, J.: Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In: Proc. ICASSP. IEEE (2015)
4. Févotte, C., Idier, J.: Algorithms for nonnegative matrix factorization with the β-divergence. vol. 23. MIT Press (2011)
5. Fu, S.W., Tsao, Y., Lu, X., Kawai, H.: Raw waveform-based speech enhancement by fully convolutional networks. In: Proc. APSIPA ASC. IEEE (2017)

6. Gao, T., Du, J., Dai, L.R., Lee, C.H.: Densely connected progressive learning for lstm-based speech enhancement. In: Proc. ICASSP. IEEE (2018)
7. Han, K., Wang, Y., Wang, D., Woods, W.S., Merks, I., Zhang, T.: Learning spectral mapping for speech dereverberation and denoising. vol. 23. IEEE (2015)
8. Hu, Y., Loizou, P.C.: Evaluation of objective quality measures for speech enhancement. IEEE TASLP **16**(1) (2007)
9. Kinoshita, K., Delcroix, M., Gannot, S., Habets, E.A., Haeb-Umbach, R., Kellermann, W., Leutnant, V., Maas, R., Nakatani, T., Raj, B., et al.: A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research. EURASIP J ADV SIG PR **2016**(1) (2016)
10. Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Habets, E., Haeb-Umbach, R., Leutnant, V., Sehr, A., Kellermann, W., Maas, R., et al.: The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In: Proc. WASPAA. IEEE (2013)
11. Li, J., Deng, L., Gong, Y., Haeb-Umbach, R.: An overview of noise-robust automatic speech recognition. IEEE/ACM TASLP **22**(4) (2014)
12. Martin, R.: Speech enhancement based on minimum mean-square error estimation and supergaussian priors. IEEE TASLP **13**(5) (2005)
13. Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., Juang, B.H.: Speech dereverberation based on variance-normalized delayed linear prediction. IEEE TASLP **18**(7) (2010)
14. Park, S.R., Lee, J.: A fully convolutional neural network for speech enhancement. arXiv preprint arXiv:1609.07132 (2016)
15. Pascual, S., Bonafonte, A., Serra, J.: Segan: Speech enhancement generative adversarial network. arXiv preprint arXiv:1703.09452 (2017)
16. Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P.: Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In: Proc. ICASSP. vol. 2. IEEE (2001)
17. Robinson, T., Fransen, J., Pye, D., Foote, J., Renals, S.: Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition. In: Proc. ICASSP. vol. 1. IEEE (1995)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proc. MICCAI. Springer (2015)
19. Takeuchi, D., Yatabe, K., Koizumi, Y., Oikawa, Y., Harada, N.: Data-driven design of perfect reconstruction filterbank for dnn-based sound source enhancement. In: Proc. ICASSP. IEEE (2019)
20. Tan, K., Wang, D.: Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement. In: Proc. ICASSP. IEEE (2019)
21. Tang, X., Du, J., Chai, L., Wang, Y., Wang, Q., Lee, C.H.: A lstm-based joint progressive learning framework for simultaneous speech dereverberation and denoising. In: Proc. APSIPA ASC. IEEE (2019)
22. Wang, K., Zhang, J., Sun, S., Wang, Y., Xiang, F., Xie, L.: Investigating generative adversarial networks based speech dereverberation for robust speech recognition. arXiv preprint arXiv:1803.10132 (2018)
23. Weninger, F., Watanabe, S., Tachioka, Y., Schuller, B.: Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition. In: Proc. ICASSP. IEEE (2014)
24. Xu, Y., Du, J., Dai, L.R., Lee, C.H.: An experimental study on speech enhancement based on deep neural networks. IEEE Signal processing letters **21**(1) (2013)
25. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Proc. DLMIA. Springer (2018)