# A Late Reverberation Power Spectral Density Aware Approach to Speech Dereverberation Based on Deep Neural Networks

Yuanlei Qi*† and Feiran Yang* and Jun Yang*†
* Key Laboratory of Noise and Vibration Research, Institute of Acoustics,
Chinese Academy of Sciences, Beijing, China
E-mail: feiran@mail.ioa.ac.cn
† University of Chinese Academy of Sciences, Beijing, China
E-mail: jyang@mail.ioa.ac.cn

***Abstract*—In recent years, a variety of speech dereverberation algorithms based on deep neural network (DNN) have been proposed. These algorithms usually adopt anechoic speech as their target output. Consequently, speech distortion might occur which impairs the speech intelligibility. As a matter of fact, early reflections can increase the strength of the direct-path sound and therefore have a positive impact on the speech intelligibility. In traditional speech dereverberation methods, early reflections are generally remained together with the direct-path sound. Based on these observations, we propose to adopt both direct-path sound and early reflections as the target DNN output in this paper. Moreover, we propose a late reverberation power spectral density (PSD) aware training strategy to further suppress the late reverberation. Experimental results demonstrate that the proposed DNN framework achieves significant improvement in objective measures even under mismatched conditions.**

## I. Introduction

The quality of speech recorded in an enclosed space is often degraded by the reflections from walls, ceilings and other objects in the room. The first 40–80 ms of the room impulse response (RIR) are generally regarded as early reflections, and reflections that arrive after the early reflections are called late reflections [1]. The combination of the direct-path sound and early reflections is referred to as the early sound component. Early reflections are actually perceived to reinforce the direct-path sound and are therefore considered useful to the speech intelligibility [2].

In recent years, many methods for speech dereverberation have been proposed. The most direct method for speech dereverberation is supposed to be bind system identification [3] and inversion. The multiple-input/output inverse theorem (MINT) [4] was the first such multi-channel inversion method. However, the MINT approach was shown to be sensitive to system estimation errors. Spectral subtraction methods [5] originally proposed for speech enhancement was also adopted to remove the late reverberation. Microphone array processing techniques such as the delay-and-sum beamformer (DSB) and its variants [6] have shown a satisfying performance for speech dereverberation especially when used for a joint reverberation and noise reduction. Recently, the weighted prediction error (WPE) algorithm [7], [8] has attracted a considerable amount of research attention. They perform multi-channel linear prediction (MCLP) at each frequency bin in the short-time Fourier transform (STFT) domain.

Recently, DNNs have become a major research subject due to their strong regression capabilities [9], [10]. In [11], it was proposed to address both dereverberation and denoising using a nonlinear DNN-based regression model. In [12], it was proposed to adopt a linear activation function at the output layer and to globally normalize the target features into zero mean and unit variance. In [13], the effects of time and frequency sampling on STFT used for speech dereverberation based on DNNs were investigated. In [14], the effects of frame shift size and context window size at the DNN input on speech dereverberation were further investigated. They proposed to estimate the reverberation time first to better select the frame shift and context window sizes for the feature extraction.

In the existing speech dereverberation algorithms based on DNNs, they attempt to adopt anechoic speech as the target output such that the DNN learns to eliminate all the reflections. As a consequence, speech distortion might occur which impairs the speech intelligibility. In many traditional dereverberation algorithms [15], it was proposed to suppress late reflections but recover the early sound component. Inspired by the traditional dereverberation algorithms, we propose to adopt both direct-path sound and early reflections as the target output in this paper. This means the DNN is trained to only suppress the late reflections. To enable this late reverberation awareness, the DNN is fed with the reverberant speech samples augmented with the late reverberation power spectral density (LRPSD). In this way, the DNN can use additional on-line late reverberation feature to better predict the early sound component. Simulation results demonstrate that the proposed DNN system achieves significant improvement in objective metrics including unseen speakers and RIRs over the baseline algorithm.

This work was supported by National Natural Science Fund of China under Grants 61501449 and 11674348, Youth Innovation Promotion Association of Chinese Academy of Sciences under Grant 2018027, IACAS Young Elite Researcher Project QNYC201812, and the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDC02020400.

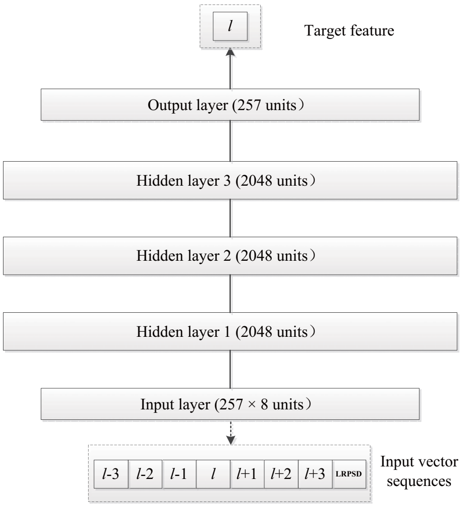
l
Target feature
Output layer (257 units)
Hidden layer 3 (2048 units)
Hidden layer 2 (2048 units)
Hidden layer 1 (2048 units)
Input layer (257 × 8 units)
l-3
l-2
l-1
l
l+1
l+2
l+3
LRPSD
Input vector sequences

Fig. 1. The architecture of the DNN for speech dereverberation.

## II. DNN-based Speech Dereverberation

The architecture of the DNN for speech dereverberation is presented in Fig. 1. The DNN in this paper includes three hidden layers, and each hidden layer includes 2048 hidden units. We use 512-point FFT and therefore the number of frequency bins is 257. In the training stage, the DNN is fed with pairs of reverberant and anechoic speech coefficients represented by the log PSD which constitute of the input and the output of the DNN in Fig. 1. The DNN is trained to map the reverberant speech coefficients to the desired anechoic speech coefficients. In the dereverberation stage, the well-trained DNN model is fed with the log PSD features of reverberant speech to generate the corresponding enhanced log PSD features of the anechoic speech. Unlike the traditional DNN, the target DNN output in this paper is the enhanced features of the early sound component, namely the combination of the direct-path sound and early reflections.

The enhanced speech signal is re-synthesized using the enhanced log PSD produced by the DNN and the phase of the original reverberant speech signal. Because human ears are shown to be insensitive to small phase distortions. In addition, it is important to use speech coefficient vectors as the input of the DNN instead of a single frame. On one hand, this can help provide more acoustic context. On the other hand, the current frame is affected by previous frames due to the effect of reverberation.

## III. LRPSD-Aware Training

In a reverberant environment, the microphone signal $x(n)$ can be denoted as the convolution of the source signal $s(n)$ and the RIR $h(m)$,

$$x(n) = \sum_{m=0}^{L_h-1} h(m)s(n-m) \tag{1}$$

where $n$ denotes the time index, and $L_h$ is the length of the RIR. The RIR can be divided into the early part $h_E(m)$ and the late part $h_L(m)$ as [15]

$$h(m) = \begin{cases} h_E(m), & 0 \le m < N_e \\ h_L(m), & N_e \le m < L_h \\ 0, & \text{Otherwise} \end{cases} \tag{2}$$

where $N_e = f_s T_{early}$ is the length of early reflections ($f_s$ being the sampling frequency) and $T_{early}$ as the early speech duration. Substituting (2) into (1), we obtain

$$x(n) = \underbrace{\sum_{m=0}^{N_e-1} h_E(m)s(n-m)}_{x_E(n)} + \underbrace{\sum_{m=N_e}^{L_h-1} h_L(m)s(n-m)}_{x_L(n)} \tag{3}$$

where $x_E(n)$ and $x_L(n)$ denote the early and late reverberant components of the speech, respectively. Transforming (3) into the STFT domain, we get

$$X(k,l) = \sum_{n=0}^{K-1} x(n+lP)w(n)e^{-j\frac{2\pi k}{K}n} = X_E(k,l)+X_L(k,l) \tag{4}$$

where $k \in \{0,1,\ldots,K-1\}$ denotes the frequency bin index, $K$ is the number of total frequency bins, $l$ denotes the frame index, $w(n)$ is the analysis window, $P$ denotes the hop size, and $X_E(k,l)$ and $X_L(k,l)$ denote the early and late reverberant components in the STFT domain.

To the best of our knowledge, the reverberation feature of each utterance was not specifically utilized in the existing speech dereverberation algorithms based on DNNs. To enable this reverberation awareness, the DNN is fed with the reverberant speech coefficients augmented with an estimate of the LRPSD. As a result, the DNN can use additional on-line late reverberation spectrum information to better suppress the late reverberation part. Therefore, the input feature vector of DNN $\mathbf{V}(k,l)$ can be written as

$$\begin{aligned} \mathbf{V}(k,l) = [&\sigma_X^2(k,l-\tau),\ldots,\sigma_X^2(k,l-1),\sigma_X^2(k,l), \\ &\sigma_X^2(k,l+1),\ldots,\sigma_X^2(k,l+\tau),\sigma_{X_L}^2(k,l)] \end{aligned} \tag{5}$$

where $\sigma_X^2(k,l)$ denotes the log PSD of $X(k,l)$, $\sigma_{X_L}^2(k,l)$ denotes the log LRPSD of late reverberation. $\tau$ denotes the number of frame expansion, and it was set to 3 in this paper which means we use 7 frames of input feature expansion except of the LRPSD frame as depicted in Fig. 1.

In recent years, a variety of LRPSD estimators have been proposed. In [6], it was shown that the proposed algorithm in [16] was simple to implementation and resulted in less speech distortion. Therefore, we choose the estimator proposed in [16] to obtain the LRPSD. In addition, we assume the reverberation time $T_{60}$ to be known to introduce fewer estimation errors. In [16], they proposed to model the RIR by an exponentially decaying random process per frequency band. Based on this model, a recursive scheme for the LRPSD estimator is given as follows

$$\hat{\sigma}_X^2(k,l) = [1-\beta]\hat{\sigma}_X^2(k,l-1) + \beta|X(k,l)|^2 \tag{6}$$

$$\begin{aligned} \hat{\sigma}_{X_R}^2(k,l) = {} & [1-\kappa]e^{-2\alpha P}\hat{\sigma}_{X_R}^2(k,l-1) \\ & + \kappa e^{-2\alpha P}\hat{\sigma}_X^2(k,l-1) \end{aligned} \tag{7}$$

$$\hat{\sigma}^2_{X_L}(k,l) = e^{-2\alpha P(N_E-1)}\hat{\sigma}^2_{X_R}(k,l-N_E+1) \tag{8}$$

where $\hat{\sigma}^2_{X_R}(k,l)$ denotes the reverberant PSD except the direct-path term, $\hat{\cdot}$ denotes estimated values, $\alpha$ is defined as $3\log 10/(f_s T_{60})$, $N_E = N_e/P$, $\beta$ is the smoothing parameter, and $\kappa$ is the shape parameter.

## IV. Experimental Results And Analysis

In our experiments, the RIRs are constructed with the image method [17]. The room dimensions are [6, 4, 3] m, and the source and microphone position are [2, 3, 1.5] m and [4, 1, 2] m, respectively. Ten RIRs were simulated with RT60 ranging from 0.1 to 1.0 s (with an increment of 0.1 s) which were convolved with all 4620 training utterances from the TIMIT set to build a large training set. Nineteen RIRs were simulated with RT60 ranging from 0.1 to 1.0 s (with an increment of 0.05 s), and they were convolved with 100 randomly selected utterances from the TIMIT test set to construct the test set.

We used a sampling frequency of 16 kHz. The frame length was 32 ms with 50% overlap. The smoothing parameter and the shape parameter were set to 0.5 and 0.8, respectively. The desired output for DNN was defined as the clean speech signals convolved with the first part of the RIRs containing early reflections until 48 ms after the direct-path sound. The perceptual evaluation of speech quality (PESQ) and the frequency weighted segmental signal-to-noise ratio (fwSegSNR) were used to measure the dereverberation performance. It is worthwhile to note that both of them are obtained by comparing the enhanced speech with the target speech, namely the early sound component.

Keras was used to train DNNs. We omit the pre-training here since large training set is available. The learning rate and the number of epochs were set to 0.00003 and 30, respectively. The mini-batch size was set to 128. The baseline algorithm in our paper is the DNN framework proposed in [12] (without pre-training), but the difference is that the desired output has been replaced by the early sound component.

### A. Spectrograms

The spectrograms of a test utterance labeled "A" at RT60 = 0.6 s were shown in Fig. 2. The DNN without LRPSD augmented (the baseline DNN, see Fig. 2(d)) achieved a PESQ increase of 0.63 compared with the unprocessed reverberant speech and removed most of the reverberation disturbance. With the LRPSD augmented (see Fig. 2(e) and Fig. 2(f)), we obtained much higher PESQ scores than the baseline DNN system which indicated the improvement of speech quality. Moreover, we observed that the low frequency contents were better restored in the spectrograms of the two proposed DNN systems compared with the baseline system.

### B. PESQ and fwSegSNR

In the followings, "Rev" denotes the unprocessed reverberant speech, "DNN-baseline" denotes the baseline DNN system, "DNN-proposed-estimate" denotes the proposed DNN system with estimated LRPSD augmented, and "DNN-proposed-oracle" denotes the proposed DNN system with oracle LRPSD augmented.

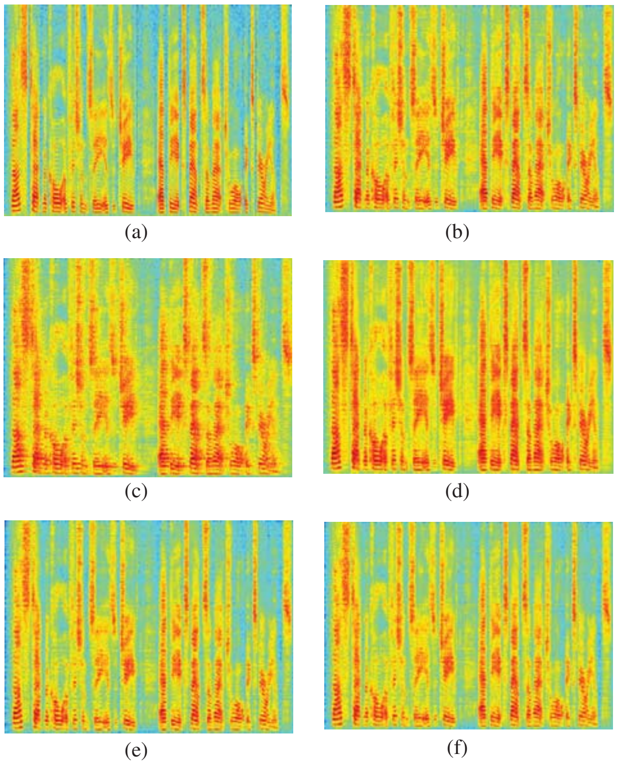

Fig. 2. Spectrograms of test utterance A, at RT60 = 0.6 s. (a) clean speech, (b) early speech (PESQ = 4.50), (c) reverberant speech (PESQ = 2.60), (d) processed by the baseline DNN, (PESQ = 3.23), (e) processed by the proposed DNN with estimated LRPSD (PESQ = 3.34), (f) processed by the proposed DNN with oracle LRPSD (PESQ = 3.49).

The average PESQ and fwSegSNR results of all DNN outputs on the test set at different RT60s were illustrated in Fig. 3 and Fig. 4 respectively. The proposed DNN system with LRPSD augmented (either estimated or oracle LRPSD) yielded higher objective measure scores than the unprocessed reverberant speech and the baseline DNN system at most RT60s, which indicated the advantage of LRPSD augmentation. Therefore, we can conclude that the augmented LRPSD can provide more on-line late reverberation spectrum information for the DNN system and thus significantly reduce the reverberation disturbance.

In addition, it was observed that the proposed DNN system with oracle LRPSD augmented achieved the best performances compared with all the other DNN systems. Specifically, the proposed DNN system with oracle LRPSD achieved a PESQ improvement of 0.77 and 0.25, respectively, compared with the unprocessed reverberant speech and the baseline DNN results. As for the fwSegSNR results, the proposed DNN system with oracle LRPSD augmented boosted the fwSegSNR by 2.1 dB and 1 dB, respectively, compared with the unprocessed reverberant speech and the baseline DNN results. The improvements of objective measure scores indicate the improvement of speech quality of the enhanced signals. It can be seen that

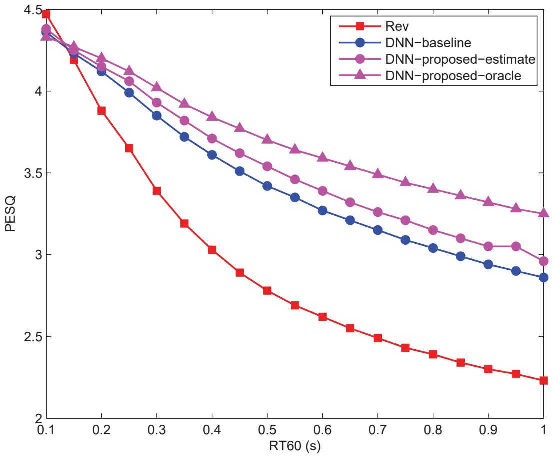

Fig. 3. Average PESQ results of the baseline and the proposed DNNs on the test set at different RT60s.

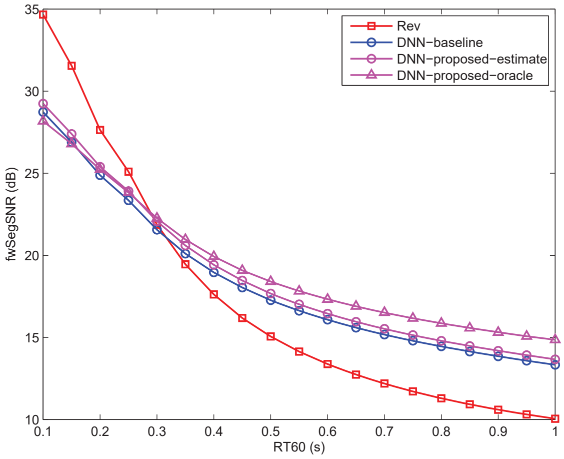

Fig. 4. Average fwSegSNR results of the baseline and the proposed DNNs on the test set at different RT60s.

the accuracy of the LRPSD estimation has a great influence on the dereverberation performance. This indicates that it is necessary to utilize more accurate LRPSD estimators in order to further improve the DNN dereverberation performance.

Simulation results demonstrated that the proposed DNN systems (“DNN-proposed-estimate” and “DNN-proposed-oracle”) achieved better performance, in particular, when the reverberation becomes relatively strong. However, when compared with the unprocessed reverberant speech at low RT60s below 0.3 s, the fwSegSNR results of all simulated DNN systems started to decrease. It would be interesting to investigate this topic in the future.

## V. Conclusions

We have proposed a new DNN framework for speech dereverberation in this paper. Inspired by traditional speech dereverberation algorithms, we propose to adopt both direct-path sound and early reflections as the target DNN output to train the model in this paper. To further suppress the late reverberant speech, the DNN is fed with the reverberant speech feature vectors augmented with an estimate of the LRPSD. With a large training set, the proposed DNN system achieved a significant improvement in terms of PESQ and fwSegSNR compared with the unprocessed reverberant speech, including mismatched conditions of RIRs and unseen speakers. The proposed DNN framework achieved a better performance especially when the reverberation is quite severe compared with the baseline DNN results.

## References

[1] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, Springer-Verlag, 2010.

[2] J. S. Bradley, H. Sato, and M. Picard, “On the importance of early reflections for speech in rooms,” *J. Acoust. Soc. Amer.*, vol. 113, no. 6, pp. 3233–3244, Jun. 2003.

[3] Y. Huang and J. Benesty, “A class of frequency-domain adaptive approaches to blind multichannel identification,” *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[4] M. Miyoshi and Y. Kaneda, “Inverse filtering of room acoustics,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[5] E. A. P. Habets, “Multi-channel speech dereverberation based on a statistical model of late reverberation,” in *Proc. IEEE ICASSP*, May 2005, pp. iv/173–iv/176.

[6] S. Braun, A. Kuklasiński, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, “Evaluation and comparison of late reverberation power spectral density estimators,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1056–1071, Jun. 2018.

[7] T. Yoshioka and T. Nakatani, “Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[8] A. Cohen, G. Stemmer, S. Ingalsuo, and S. Markovich-Golan, “Combined weighted prediction error and minimum variance distortionless response for dereverberation,” in *Proc. IEEE ICASSP*, Mar. 2017, pp. 446–450.

[9] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, “Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation,” *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, pp. 1–18, Dec. 2016.

[10] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[11] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, “Learning spectral mapping for speech dereverberation and denoising,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 982–992, Jun. 2015.

[12] B. Wu, K. Li, M. L. Yang, and C.-H. Lee, “A study on target feature activation and normalization and their impacts on the performance of DNN based speech dereverberation systems,” in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Dec. 2016.

[13] B. Wu, K. Li, M. L. Yang, and C.-H. Lee, “A study on sampling of STFT modifications in time and frequency domains for DNN-based speech dereverberation,” in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Dec. 2016.

[14] B. Wu, K. Li, M. L. Yang, and C.-H. Lee, “A reverberation-time-aware approach to speech dereverberation based on deep neural networks,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 102–111, Jan. 2017.

[15] M. Parchami, W. P. Zhu, and B. Champagne, “Model-based estimation of late reverberant spectral variance using modified weighted prediction error method,” *Speech Commun.*, vol. 92, pp. 100–113, Sep. 2017.

[16] E. A. P. Habets, S. Gannot, and I. Cohen, “Late reverberant spectral variance estimation based on a statistical model,” *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–774, Sep. 2009.

[17] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.