LATE REVERBERATION SUPPRESSION USING RECURRENT NEURAL NETWORKS WITH LONG SHORT-TERM MEMORY

Yan Zhao[1] *DeLiang Wang*[1,2,3] *Buye Xu*[4] *Tao Zhang*[4]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA
[3]Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, China
[4]Starkey Hearing Technologies, USA

{zhao.836, wang.77}@osu.edu, {buye_xu, tao_zhang}@starkey.com

ABSTRACT

Human speech is usually distorted by room reverberation. These corruptions degrade speech quality and intelligibility, especially under a long reverberation time, and they also pose a serious problem for many speech-related applications such as automatic speech recognition. In this paper, we propose a supervised speech dereverberation algorithm that models late reverberation using a recurrent neural network (RNN) with long short-term memory (LSTM). By taking advantage of LSTM's ability to capture a long history, late reverberation can be effectively removed by the proposed approach. Systematic evaluations indicate that our approach improves the quality of reverberant speech in a wide range of reverberant conditions. Moreover, the proposed system is a causal system, which can be applied in real-time applications.

Index Terms— speech dereverberation, long short-term memory, recurrent neural networks, supervised speech enhancement

1. INTRODUCTION

Reverberation is an acoustic phenomenon caused by the reflections of sound waves in a room. Strong reverberation can significantly degrade speech intelligibility and sound quality for human listeners [1]. It has been shown that the early part of the reverberation, the reflections arriving within 50 ms after the direct sound, is actually beneficial for intelligibility. It is the late reverberation, which arrives 50 ms after the direct sound, degrades the speech intelligibility [2]. Reverberation can also impose challenges for machine systems involving speech processing. For example, the performance of far-field automatic speech recognition (ASR) systems often suffer due to the existence of the room reverberation [3]. Reducing the effect of reverberation has been an important topic of speech processing.

This work is supported in part by Starkey, NIDCD grants (R01 DC012048 and R01 DC015521), and the Ohio Supercomputer Center.

Many single-channel dereverberation algorithms have been proposed in the past decays. Lebart *et al.*, for example, propose a spectral subtraction algorithm to suppress late reverberation by assuming an exponential decay model of reverberation [4]. Wu and Wang propose a two-stage algorithm which cancels early reflections with an inverse filter and reduces late reverberation based on spectral subtraction [5]. Long-term linear prediction based dereverberation methods [6, 7] have been shown to be very effective for late reverberation suppression. In these algorithms, the frequency-dependent linear prediction filters are first obtained based on a number of history frames using the weighted prediction error (WPE) minimization. The enhanced signal is then obtained by subtracting the filtered signal from the original reverberant signal in the subband domain. López *et al*. [8] propose to estimate the amplitude of late reverberation by modeling it as a sparse linear combination of the amplitudes of past frames, and then solve the problem using Lasso. Good dereverberation performance is achieved.

Recently, many supervised speech enhancement algorithms are proposed to perform dereverberation and have achieved substantial improvements over the traditional methods. In [9], Han *et al*. propose to learn a spectral mapping function from the log magnitude spectrum of reverberant speech to that of anechoic speech by using a deep neural network (DNN). Wu *et al*. [10] point out the importance of reverberation time dependent parameters during training a DNN-based dereverberation system. Then they propose a reverberation-time-aware approach to remove reverberation, which outperforms Han *et al*.'s approach. In Weninger *et al*.'s robust speech recognition system [11], dereverberation is performed by using a deep bi-directional recurrent neural network (RNN) with long short-term memory (LSTM). Taking the phase into account, Williamson and Wang [12] propose to estimate a complex ideal ratio mask (cIRM) by using a DNN to remove reverberation in the complex domain. As the joint enhancement of magnitude and phase spectrum, better results are reported in their study. Comparing to the

traditional methods mentioned above, these supervised methods do not explicitly estimate either the early reflections or the late reverberation, but attempt to directly construct the dry speech from the reverberant observation.

One critical limitation of those supervised algorithms, however, is that they are non-causal since the information from the future is used in the processing. In this paper, we propose a causal supervised dereverberation algorithm. Instead of deriving the dry speech directly, a uni-directional LSTM RNN [13] is utilized to estimate late reverberation, which is then subtracted from the original signal to enhance the reverberant speech.

The rest of the paper is organized as follows. We will describe our proposed algorithm in details in the next section. The experimental setup and evaluation results are presented in Section 3 and Section 4, respectively. We conclude this paper in Section 5.

2. ALGORITHM DESCRIPTION

2.1. Problem formulation

Let $s(t)$ and $h(t)$ denote anechoic speech and room impulse response (RIR), respectively. The reverberant speech $y(t)$ is modelled by

$$y(t) = s(t) * h(t) \tag{1}$$

where $*$ denotes a convolution operation. According to the arrival time of the signal, the reverberant speech can be divided into two components, namely, direct sound plus early reflections and late reverberation. By writing the RIR into two portions, h_{de} and h_l, the reverberant speech can be represented by

$$y(t) = s(t) * h_{de}(t) + s(t) * h_l = y_{de}(t) + y_l(t) \tag{2}$$

The objective of this study is to remove the late reverberation component $y_l(t)$ from the corresponding reverberant speech $y(t)$.

2.2. Features and training target

Given a time-domain signal, we divide it into frames by using a 32 ms Hamming window with 8 ms window shift. 512-point fast Fourier transformation (FFT) is applied to each frame, which results in 257 frequency bins. In our study, the magnitude spectrum of the reverberant speech is directly used as features. To compress the dynamic range of the values, a cubic root compression is applied. All the features are normalized to zero mean and unit variance by using the statistics of the training data. We use $\boldsymbol{Y}(m)$ to denote the normalized compressed magnitude features at time frame m, which is a 257-dimension vector. Then, our proposed system takes the following sequential feature vectors as the input,

$$\boldsymbol{Y} = \{\boldsymbol{Y}(1), \boldsymbol{Y}(2), ..., \boldsymbol{Y}(N)\} \tag{3}$$

where N is the total number of frames in the utterance. At each time step, the features of one frame are fed to the system. In other words, no context window is employed.

As for the training target of the proposed system, at time step t, one choice is to use the late reverberation in the corresponding time frame as the target. During testing, after obtaining the estimation of late reverberation, we can utilize techniques like spectral subtraction to remove the late reverberation part. In this paper, however, the magnitude spectrum of the direct sound plus early reflections is used as the training target to simplify the processing. Similar to the input features, a cubic root operation is also applied to compress the values. Let $\boldsymbol{Y}_{de}(m)$ denote the compressed magnitude spectrum of the direct sound plus early reflections at time frame m. The training target can be expressed by the following sequential vectors,

$$\boldsymbol{Y}_{de} = \{\boldsymbol{Y}_{de}(1), \boldsymbol{Y}_{de}(2), ..., \boldsymbol{Y}_{de}(N)\} \tag{4}$$

It should be pointed out that different from the input features, we do not perform mean and variance normalization on the training target.

2.3. Network architecture

It is necessary to exploit long-term history information when we perform dereverberation. With the internal memory, RNN is designed to model sequential data with long-term dependencies. However, gradient vanishing and exploding issues make a vanilla RNN hard to optimize. By introducing a memory cell and employing gate mechanism to control the information flow, LSTM RNN has shown powerful ability to model long range dependencies embedded in the sequential data. Consequently, in order to capture the long history in the past observed reverberant speech, we utilize LSTM RNN to predict late reverberation. The LSTM block used in this study is defined by the following equations,

$$\begin{aligned} i_t &= sigmoid(W_{ii}x_t + W_{hi}h_{t-1} + b_i) \\ f_t &= sigmoid(W_{if}x_t + W_{hf}h_{t-1} + b_f) \\ g_t &= tanh(W_{ig}x_t + W_{hg}h_{t-1} + b_g) \\ o_t &= sigmoid(W_{io}x_t + W_{ho}h_{t-1} + b_o) \\ c_t &= f_t \circ c_{t-1} + i_t \circ g_t \\ h_t &= o_t \circ tanh(c_t) \end{aligned} \tag{5}$$

where i_t, f_t, g_t, o_t are the input, forget, cell, and output gates, respectively; at time step t, h_t is the hidden state; c_t is the memory cell state; x_t is the input of the first layer or the hidden state of the previous layer; $W's$, $b's$ denote the weights and biases in the linear transformations, respectively; $\circ$ denotes the element-wise multiplication.

Fig. 1 shows the system diagram of the proposed algorithm. We present three time steps of the system for better understanding. The input features of each time step are directly fed to the LSTM RNN with two hidden layers. A linear

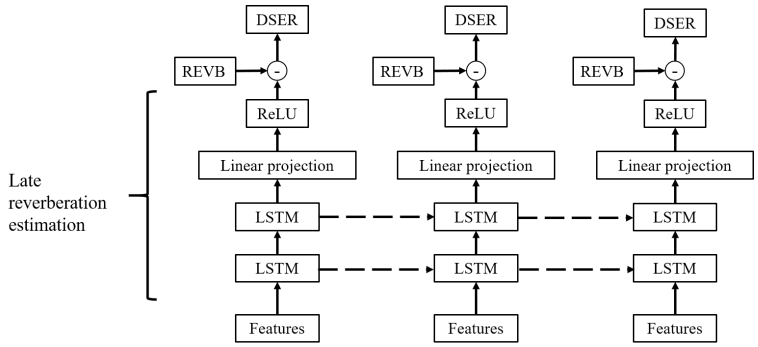


Fig. 1: System diagram of the proposed algorithm. "REVB" denotes the compressed magnitude spectrum of the reverberant speech; "DSER" denotes the compressed magnitude spectrum of the direct sound plus early reflections.

layer is on top of the LSTM RNN to project the hidden states of the last layer of LSTM RNN to the late reverberation. We use rectified linear units (ReLU) [14] after the linear projection layer to guarantee the positive estimation of the late reverberation. Then, we subtract the late reverberation prediction from the magnitude spectrum of the reverberant speech to obtain the magnitude spectrum of the direct sound plus early reflections as the system output. It should be pointed out that the values of the magnitude spectra are compressed by cubic root function. In other words, the spectral subtraction is performed in a cubic root compressed space. Although we are not explicitly using the late reverberation as the training target, the system forces the LSTM RNN to learn to estimate the late reverberation. If we only look at the input and the output of the proposed system and consider it as a black box, it achieves a frame-level sequence mapping from the magnitude spectrum of the reverberant speech to that of the direct sound plus early reflections, which is similar to perform a sequential spectral mapping. However, the internal mechanism is totally different from the spectral mapping approach and no context window is needed.

3. EXPERIMENTAL SETUP

The proposed system is evaluated by using the IEEE corpus [15] spoken by a female speaker. There are 72 phonetically balanced lists of sentences in the corpus, with 10 sentences in each list. In our experiments, we select sentences from List 1-50, List 67-72 and List 51-60 to construct training data, validation data and test data, respectively. We simulate a reverberant room of size $10\ m \times 7\ m \times 3\ m$. Different RIRs are generated by placing an omnidirectional microphone at a fixed position while randomly choosing the position of the speaker. In addition, the distance between the receiver (microphone) and the speaker is set to $2\ m$. An RIR generator [16] is utilized to produce different RIRs in the room, which employs the image method [17]. In the study, a wide range of reverberation times are investigated, from 0.3 s to 1.0 s, with an increment of 0.1 s. Under each T_{60}, 10 different RIRs are generated for training and validation; 1 RIR is generated for testing. As a consequence, there are 500×8 (T_{60}s)×10 (RIRs) = 40,000 reverberant utterances in the training set; 50×8 (T_{60}s)×10 (RIRs) = 4,000 reverberant utterances in the validation set; 100×8 (T_{60}s)×1 (RIR) = 800 reverberant utterances in the test set. It should be pointed out that both the RIRs and sentences used for testing are unseen during training/validation. We denote this test set as "Test A".

In order to test whether the trained model can generalize to other rooms, another set of RIRs are also generated for testing. Three RIRs with the values of T_{60} at 0.3 s, 0.6 s and 0.9 s are generated in a reverberant room of size $8\ m \times 9\ m \times 2.5\ m$. The distance between the receiver and the speaker is $1.8\ m$. Therefore, we have another test set with 100×3 (T_{60}s)×1 (RIR) = 300 reverberant utterances to process. This test set is denoted by "Test B".

In the proposed system, we are using a LSTM RNN with two hidden layers and 512 units in each layer. In the experiments, we find that stacking more LSTM layers can only bring very slightly performance improvements for our task. Considering the system complexity and the size of training data, two LSTM layers are employed in the current system. To avoid overfitting issue during training, we apply dropout with 0.3 dropout rate between the two hidden layers of the LSTM RNN [18]. Furthermore, we also utilize the weight-dropped technique [19] to mitigate overfitting across the recurrent connections. The dropout rate for weight-dropped LSTM is set to 0.5. Instead of truncating the utterance with fixed length sequence, we use a whole utterance as a sequence and perform batch processing with variable length sequences. The batch size is 8 in our experiments. The parameters are initialized by using orthogonal initialization method [20]. We train the system by using Adam [21] optimizer and mean squared error (MSE) loss. The time-domain signal is resynthesized by using the reverberant phase.

	PESQ									SNR_{fw} (dB)								
T_{60} (s)	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	**Avg.**	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	**Avg.**
unprocessed	2.911	2.708	2.346	2.296	2.098	1.996	1.957	1.825	2.267	20.30	15.67	10.87	9.35	7.83	6.63	5.58	4.82	10.13
WPE	3.032	2.941	2.541	2.601	2.302	2.111	2.088	1.906	2.440	15.91	15.59	14.38	13.59	11.68	9.59	8.69	7.49	12.12
López *et al.*	2.733	2.762	2.485	2.531	2.351	2.317	2.331	2.136	2.456	6.39	6.56	6.14	6.25	6.10	6.15	5.96	5.60	6.14
proposed	**3.522**	**3.390**	**3.063**	**3.036**	**2.849**	**2.696**	**2.711**	**2.546**	**2.977**	**21.14**	**18.46**	**16.72**	**16.01**	**15.35**	**13.39**	**13.55**	**12.72**	**15.92**

Table 1: Average PESQ and SNR_{fw} scores on Test A. Boldface number indicates the best performance.

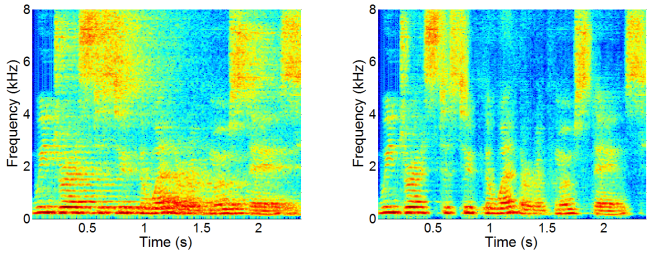
	PESQ				SNR_{fw} (dB)			
T_{60} (s)	0.3	0.6	0.9	**Avg.**	0.3	0.6	0.9	**Avg.**
unprocessed	2.760	2.230	2.004	2.331	18.06	9.00	5.49	10.85
WPE	2.842	2.542	2.094	2.493	15.59	13.05	8.66	12.43
López *et al.*	2.697	2.527	2.323	2.516	6.57	6.30	6.00	6.29
proposed	**3.385**	**3.014**	**2.685**	**3.028**	**19.03**	**15.88**	**13.13**	**16.01**

Table 2: Average PESQ and SNR_{fw} scores on Test B. Boldface number indicates the best performance.

4. EVALUATION RESULTS

In this sudy, we utilize perceptual evaluation of speech quality (PESQ) [22] (the value range is [-0.5, 4.5]) and the frequency-weighted segmental signal-to-noise ratio (SNR_{fw}) [23] to evaluate the proposed approach. Since only late reverberation is removed, the direct sound plus early reflections is used as the reference signal. For both metrics, the higher number indicates the better performance. We compare our approach with two late reverberation suppression approaches, one is the WPE approach [1] and the other is López *et al.*'s Lasso-based dereverberation approach [2].

Table 1 and Table 2 list the average PESQ and SNR_{fw} scores on Test A set and Test B set, respectively. We firstly compare the performance of the three approaches on Test A set. In this set, a wide range of reverberant conditions are investigated. Our proposed approach shows the best dereverberation performance under all the conditions in terms of both PESQ and SNR_{fw} values. Averagely, compared with the reverberant speech, the enhanced speech improves the PESQ score around 0.7, and the SNR_{fw} nearly 6 dB, indicating the effectiveness of our approach to perform dereverberation. According to the PESQ scores, both the WPE approach and López *et al.*'s approach can provide limited improvement for the sound quality. López *et al.*'s approach becomes better when dealing with longer reverberation times in terms of the PESQ scores, however, it performs very poorly under the SNR_{fw} measurement. Similar performance trends can be observed on Test B set. As we mentioned in Section 3, the main purpose of the experiments on Test B is to answer the question whether the model trained in one room of fixed size can be generalized to other rooms. If it fails, the applications of the proposed approach will be very limited. The PESQ and SNR_{fw} scores in Table 2 demonstrate that our trained dereverberation model can generalize very well to untrained room conditions. Significant improvements over the WPE approach and López *et al.*'s approach are observed.



(a) reverberant speech (b) enhanced speech

Fig. 2: (Color online) Example spectrograms of reverberant speech (T_{60} = 1.0 s) and the corresponding enhanced speech.

To illustrate the effectiveness of the proposed system to suppress the late reverberation, one enhancement example is given in Fig. 2. The sentence is randomly chosen from Test A with the most severe reverberant condition (T_{60} is 1.0 s). The content of the selected sentence is "We dress to suit the weather of most days." Fig. 2(a) presents the spectrogram of the reverberant speech. The corresponding spectrogram of the enhanced speech processed by our approach is shown in Fig. 2(b). Obviously, most smearing effects caused by late reverberation have been removed and the speech structures in the T-F representation are recovered. This demonstrates that late reverberation can be largely suppressed by our proposed approach under adverse reverberant conditions with long reverberation time.

5. CONCLUSION

In this paper, we have proposed a LSTM based system to perform late reverberation suppression. By leveraging the capacity of recurrent connections to model the long-term dependencies in reverberant speech, late reverberation is well estimated and removed. Moreover, the causality of our system makes it possible to deploy in real-time applications. Systematic evaluations under a wide range of reverberant conditions have shown that late reverberation is better removed by the proposed approach than other related methods.

[1] The software is available at `http://www.kecl.ntt.co.jp/icl/signal/wpe/`

[2] We are using the software provided by Nicolás López in the experiments.

6. REFERENCES

[1] K. S. Helfer and L. A. Wilber, "Hearing Loss , Aging , and Speech Perception in Reverberation and Noise," *Journal of speech and hearing research*, vol. 33, no. March, pp. 149–155, 1990.

[2] A. K. Nábělek, T. R. Letowski, and F. M. Tucker, "Reverberant overlap-and self-masking in consonant identification," *The Journal of the Acoustical Society of America*, vol. 86, pp. 1259–1265, 1989.

[3] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, pp. 114–126, 2012.

[4] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[5] M. Wu and D. L. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 774–784, 2006.

[6] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1717–1731, 2010.

[7] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 2707–2720, 2012.

[8] N. López, Y. Grenier, G. Richard, and I. Bourmeyster, "Single channel reverberation suppression based on sparse linear prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5182–5186.

[9] K. Han, Y. Wang, D. L. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 982–992, 2015.

[10] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, pp. 102–111, 2017.

[11] F. Weninger, S. Watanabe, Y. Tachioka, and B. Schuller, "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4623–4627.

[12] D. Williamson and D. L. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–1780, 1997.

[14] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[15] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio Electroacoust*, vol. 17, pp. 225–246, 1969.

[16] E. Habets, "Room impulse response generator," Available at `https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator`.

[17] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[18] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

[19] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing lstm language models," *arXiv preprint arXiv:1708.02182*, 2017.

[20] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.

[21] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, vol. 2, pp. 749–752.

[23] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, pp. 3387–3405, 2009.