A RETURN TO DEREVERBERATION IN THE FREQUENCY DOMAIN USING A JOINT LEARNING APPROACH

Yuying Li†, and Donald S. Williamson‡

†Department of Intelligent Systems Engineering, Indiana University, USA
‡Department of Computer Science, Indiana University, USA
{liyuy,williads}@indiana.edu

ABSTRACT

Dereverberation is often performed in the time-frequency domain using mostly deep learning approaches. Time-frequency domain processing, however, may not be necessary when reverberation is modeled by the convolution operation. In this paper, we investigate whether deverberation can be effectively performed in the frequency-domain by estimating the complex frequency response of a room impulse response. More specifically, we develop a joint learning framework that uses frequency-domain estimates of the late reverberant response to assist with estimating the direct and early response. We systematically compare our proposed approach to recent deep learning based approaches that operate in the time-frequency domain. The results show that frequency-domain processing is in fact possible and that it often outperforms time-frequency domain based approaches under different conditions.

Index Terms— Deep neural network, frequency response, speech quality, dereverberation

1. INTRODUCTION

Reverberation degrades perceptual speech quality and intelligibility as sound reflections obscure signal structure. This creates a challenge for many applications, including, hearing aids, automatic speech recognition and speaker identification. Many methods have been proposed to remove reverberation. Zhao *et al.*, for example, proposed a long short-term memory (LSTM) deep neural network to predict late reflections in the time-frequency (T-F) domain [1]. The predicted late reverberation is subsequently subtracted from the reverberant speech signal. Han *et al.* develop a spectral mapping approach that predicts anechoic speech from reverberant speech by splicing the log-magnitude response and using a deep neural network (DNN) for prediction [2]. Williamson *el al.* use a complex ratio mask (cIRM) [3] for derverberation, where the mask enhances the magnitude and phase T-F responses [4]. The approach takes a complementary set of T-F features as inputs and estimates the cIRM, which is then used to recover anechoic speech. Nakatani *et al.* proposed a unsupervised technique, known as the weighted prediction error (WPE), which estimates an inverse filter that is subtracted from the reverberated speech [5, 6]. Delfarah *et al.* proposed a two staged mask estimation that uses multi-dimension features and estimates T-F domain masks [7]. Sun *et al.* proposed an approach that took T-F domain features as input and predicts dereverberation mask (DM) and IRM to recover the speech [8]. In [9], a recurrent neural network (RNN) based approach is used for dereverberation. The above approaches use different machine learning techniques to perform dereverberation, but the estimation process is always conducted in the time-frequency domain.

Many years ago, however, dereverberation was performed in the frequency-domain using different signal processing approaches. In [10], inversion is used to deconvolve mixtures that are generated through convolution. A two-stage inverse filtering algorithm is developed in [11], where the implementation occurs in the frequency domain. Many other frequency-domain dereverberation methods were also proposed [12, 13, 14]. These approaches use methods such as independent component analysis (ICA) that required assumptions based on the room impulse response to hold, which is not always possible [15]. Frequency-domain deep learning-based approaches, however, may not need to make these assumptions, so it may be possible to return to frequency-domain processing.

In this paper, we train a joint LSTM dereverberation approach that uses an estimate of the late reverberation transfer function to help predict the transfer function of the direct and early signal. This approach operates completely in the frequency domain. This is done because reverberation is modeled by the convolution operator, which can be performed in the frequency domain, regardless of the time-domain structure of speech. More specifically, the joint deep neural network is trained jointly using both real and imaginary components to enhance magnitude and phase information.

The rest of this paper is organized as follows. Section 2 discusses the relation to prior work. Section 3 describes the details of the problem and proposed model. Our experimental setup is given in section 4. Section 5 discusses the results and comparison approaches. Finally, a conclusion is given in Section 6.

2. RELATION TO PRIOR WORK

The work proposed here focuses on speech dereverberation in the complex frequency domain. Previous studies on this topic perform dereverberation in the spectral-magnitude domain [2, 7, 6, 16]. Although complex domain dereverberation is presented in [17, 3], their approach is mainly focus on T-F domain features. In addition, techniques in [1] predict the magnitude of late speech, and subtract that from the reverberant speech, so it does not estimate the late and direct plus early responses together. Our approach, on the other hand, requires only frequency information and utilizes both late and direct plus early information to predict the transfer function of the room impulse response (RIR).

3. ALGORITHM DESCRIPTION

3.1. Problem Description

A reverberant speech signal can be computed as the convolution of a clean speech signal $s(t)$ with a RIR, $h(t)$

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

where * denotes convolution. The RIR can be decomposed into the sum of the direct RIR, early RIR and late RIR (e.g., $h(t) = h_d(t) + h_e(t) + h_l(t)$). Replacing $h(t)$ in equation 1 with this definition, results in:

$$\begin{aligned} x(t) &= s(t) * h_d(t) + s(t) * h_e(t) + s(t) * h_t(t) \\ &= s(t) * h_{de}(t) + s(t) * h_l(t) \end{aligned} \tag{2}$$

where $h_d(t)$ denotes the RIR of the direct sound, $h_e(t)$ denotes the RIR of the early reflections, $h_l(t)$ denotes the RIR of the late reflections, and $h_{de}(t)$ denotes the RIR of the direct sound plus early reflections. The objective of this study is to remove the late reflections (e.g. $x_l(t) = s(t) * h_l(t)$) from the corresponding reverberant signal.

3.2. Features and training labels

Given a reverberant speech signal in the time domain, we compute the 1024-point discrete Fourier transform (DFT). We then concatenate the real and imaginary components of the 1024-point DFT as our input. This input is normalized using Min-Max normalization that is calculated across all real and imaginary components separately [18]. This results in values between 0 and 1. Let $Y(m)$ denote the normalized and concatenated input for the m^{th} signal. $Y(m)$ and the full input matrix into our proposed system are defined as,

$$\begin{aligned} Y(m) &= \begin{bmatrix} Y_i(1)Y_i(2)\dots Y_i(N) \\ Y_r(1)Y_r(2)\dots Y_r(N) \end{bmatrix} \\ Y &= \{Y(1)Y(2)\cdots Y(N_S)\} \end{aligned} \tag{3}$$

where we assume the frequency indexing starts at 1 and N denotes the finite number of DFT points, which is 1024 in this study, and N_S denotes the number of training samples.

We elect to predict transfer functions of the RIRs instead of speech. Recent work has shown that predicting masking outperforms predicting speech in the time-frequency domain [19], so in this case, we elect to predict the transfer function instead of the speech itself. We transform the direct plus early RIR ($h_{de}(t)$) and the late RIR ($h_l(t)$) into 1024-point DFTs. We also concatenate the real and imaginary components of the resulting DFT into one matrix, and this serves as the training label. The labels can be expressed by the following feature matrix,

$$\begin{aligned} H_l(m) &= \{H_l(1)H_l(2)\cdots H_l(N)\} \\ H_{de}(m) &= \{H_{de}(1)H_{de}(2)\cdots H_{de}(N)\} \end{aligned} \tag{4}$$

As is done for the input data, we did Min-Max normalization on the training labels as well since this improved performance. During testing, we de-normalize the output based on the prior calculated min and max values.

3.3. Network architecture

We use a joint deep neural network to estimate the frequency response of the RIRs. The joint network is trained to first map the features of the reverberant signal to the frequency response of late RIR. This estimate is then combined with the input features and supplied to another network that estimates the frequency response for the direct plus early RIR. Fig. 1 shows the network structure of the joint LSTM network. We experiment with a joint fully-connected network and a joint LSTM network.

The joint-fully connected network (joint FCN) passes the input matrix through three fully-connected hidden layers, where the first two layers have 2048 units and the third hidden layer has 1024 units. Rectified linear activation functions are used in the hidden layer. Then, we take the output of the first three hidden layers (estimate of late transfer function), and concatenate it with the original input feature matrix, and this becomes a new feature matrix. We pass this new feature matrix to three similar fully connected layers, where the first hidden layer has 4096 units since the new input matrix for this layer is double the size of the input matrix. Also, these three layers used rectified linear activation functions.

For the joint-LSTM approach, the input features are provided to three LSTM layers, as shown in Fig. 1. The hidden size is set to 2048. As is shown in Fig. 1, we only take the last hidden state from the third hidden layer, which outputs a vector of size 2048. This output is reshaped and concatenate with the original input matrix as a new input for the next stage. Then we set the hidden size to 2048 units. Finally, we take the last hidden state from the last hidden layer as our output. Rectified linear activation functions are applied to all LSTM layers.

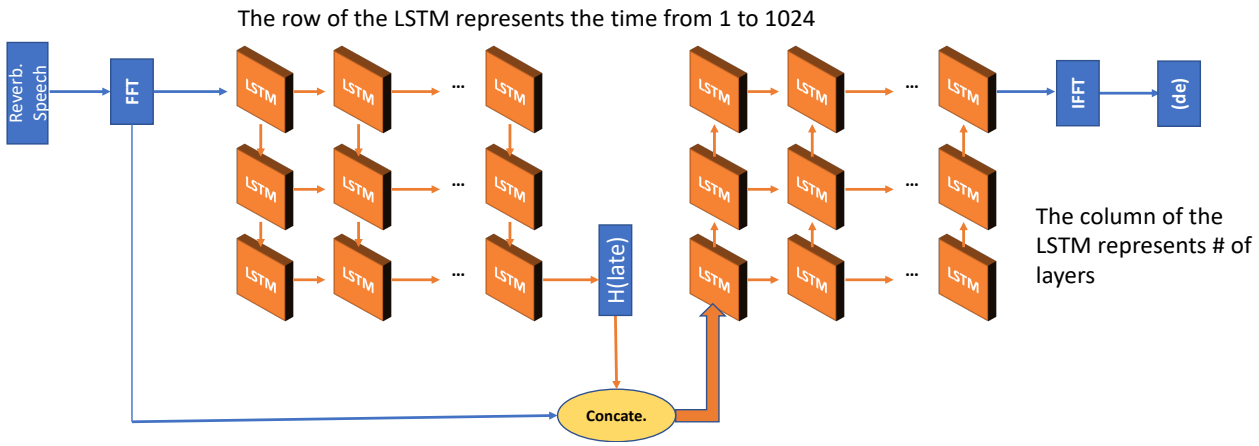

Fig. 1. Illustration of the proposed joint approach.

We train the joint-FCN and joint-LSTM networks with Adam optimizer, and the learning rate is set to 0.0001. We use the mean-square error (MSE) as the objective function. The batch size is set to 32 for joint FCN, and to 20 for joint LSTM. All signals are either truncated or padded to the mid length of the sample.

The joint FCN and joint LSTM networks are trained using the standard back propagation algorithm with mean-square error cost function,

$$\frac{1}{2N}\sum_f\left[\left(\hat{H}_l^r(f)-H_l^r(f)\right)^2+\left(\hat{H}_l^i(f)-H_l^i(f)\right)^2\right] +\frac{1}{2N}\sum_f\left[\left(\hat{H}_{de}^r(f)-H_{de}^r(f)\right)^2+\left(\hat{H}_{de}^i(f)-H_{de}^i(f)\right)^2\right] \tag{5}$$

where $\hat{H}_l^r(f)$ and $\hat{H}_l^i(f)$ are the estimated real and imaginary components of late RIR's transfer function that are generated from the first stage, N is the number of features of each input, which is assumed to be 1024 units in this study. Similarly, $\hat{H}_{de}^r(f)$ and $\hat{H}_{de}^i(f)$ are the real and imaginary components that are generated by the whole network.

The joint network estimates the normalized version of $\hat{H}_{de}^r(f)$ and $\hat{H}_{de}^i(f)$. During the testing phase, the values are de-normalized using the following:

$$\begin{aligned}\hat{H}_{de}^r(f)&=\hat{H}_{de}^r(f)\Big(\text{max(r)}-\text{min(r)}\Big)+\text{min(r)}\\ \hat{H}_{de}^i(f)&=\hat{H}_{de}^i(f)\Big(\text{max(i)}-\text{min(i)}\Big)+\text{min(i)}\end{aligned} \tag{6}$$

where max(r) and min(r) denotes the max and min value in real components of direct plus early FFT of RIRs, and max(i) and min(i) denotes the max and min value in imaginary components of direct plus early FFT of RIRs.

4. EXPERIMENTS

4.1. Experimental setup

The proposed system is evaluated using the TIMIT corpus [20], which is spoken by people from different regions of the United States. In our experiments, we randomly choose 3000, 1000, and 1000 sentences to construct our training, validation and testing datasets. All 5000 sentences are down-sampled to 16kHz. We simulate RIRs from 5 different rooms, with respective dimensions of: $9m \times 8m \times 7m$, $10m \times 7m \times 3m$, $6m \times 6m \times 10m$, $8m \times 10m \times 4m$ and $7m \times 7m \times 8m$, respectively. The distance between the receiver and the speaker is set to 1m in all cases. The image method is used to generate the RIRs [21]. We select 3 different reverberation times (i.e., 0.3s, 0.6s and 0.9s) to represent low, moderate, and high reverberation conditions. We simulated 500 different RIRs for each T_{60}, resulting in 1500 different RIRs for each room for training. 500 different RIRs are generated from the fourth room for validation data, and another 500 different RIRs are generated from the fifth room for test data. In order to make the dataset reasonable for training, we combine each RIR with two different speech signals. As a result, there are $3000\times3(T_{60}\text{s})$= 9,000 reverberant utterances in the training set; $1000\times3(T_{60}\text{s})$ = 3,000 reverberant utterances in both the validation and testing sets.

To avoid round off errors and to ensure the network learns, we scaled RIRs by 1000 before we combine it with the clean speech. The training targets are also scaled by 1000. During testing, we remove the scaling, and convert back into the time domain. Since we assume the known size of FFT is 1024, we can only inverse the FFT back to the same size. In order to evaluate the performance of our approach, we convolve the predicted direct plus early RIRs with clean speech, and compared with direct plus early speech we generated through the image method.

Table 1. Comparison with different approaches

	SDR (dB)			STOI			PESQ		
	0.3	0.6	0.9	0.3	0.6	0.9	0.3	0.6	0.9
Mixture	-1.89	-2.98	-4.01	0.58	0.48	0.43	1.7	1.42	1.25
Joint FCN	7.69	7.23	8.03	**0.77**	0.66	0.67	**2.28**	2.02	2.11
Joint LSTM	7.53	**9.32**	**8.92**	0.60	0.64	0.68	2.12	**2.10**	**2.13**
cIRM [17]	**7.90**	7.32	7.95	0.73	**0.70**	**0.71**	**2.28**	2.03	2.09
IRM [7]	7.62	7.27	7.54	0.72	**0.70**	0.70	2.23	2.00	2.01
Spectral Mapping [2]	7.28	7.25	7.48	0.71	0.69	0.68	2.01	1.97	1.90

4.2. Comparison Approaches

In our experiments, we compare our proposed approach with three baseline approaches: cIRM estimation [17], the Idea Ratio Mask (IRM) similar to [7], and Spectral Mapping approach similar to [2]. We train Spectral Mapping, IRM and cIRM with a 3-layer DNN, which is done in the original approaches. Also, we modify the target of Spectral Mapping to the direct plus early speech and re-define the IRM and cIRM oracle masks, so that the approaches are aligned with our proposed approach.

5. RESULTS

The perceptual evaluation of speech quality (PESQ) [22], Short-Time Objective intelligibility (STOI) measure and signal to distortion (SDR) are used to evaluate performance. The PESQ score ranges from -0.5 to 4.5, the STOI measure range from 0 to 1, and SDR has no specific range, but for all three, higher values indicate better performance.

Table 1 lists all the comparison scores for the different approaches. The best performance is shown in bold. In terms of SDR, all approaches get reasonable scores. cIRM particularly performs best at T_{60} of 0.3. Our joint-LSTM approach outperforms the comparison approaches as the T_{60} increases. In terms of STOI score, for the proposed approaches, there are some improvements compared to the mixture, but not as good as cIRM and IRM, which perform better when T_{60} increases. When T_{60} is 0.3, our Joint-FCN approach outperforms the others. The best STOI score is 0.19 higher when T_{60} is 0.3, around 0.28 higher then when T_{60} ia 0.6 and 0.9. We also observed that our Joint LSTM outperforms the baseline approaches by 0.07 in PESQ score when T_{60} is 0.6, and 0.06 higher when T_{60} at 0.9 compared to cIRM. Compared to the mixture, the PESQ score of our joint FCN increases by roughly 0.5 when T_{60} is 0.3 and even higher at the other two cases. Overall, our joint learning approaches perform better as T_{60} increases, which is a great sign to the dereverberation area since most of the time the approach will get worse results when T_{60} increases. More noticeably, the results reveal that frequency-domain processing often outperforms T-F domain processing. Indicating that this form of dereverberation is promising.

6. CONCLUSION

In this paper, we proposed a deep learning approach to extract frequency information of RIRs out solely from frequency information of reverberant speech. This approach manages to extract RIR frequency information and enhance the reverberant speech. This approach enhances reverberant speech in complex domain. Our approach deviates from recent methods by processing in the frequency domain. In addition, the joint deep neural network method by adding predicted late information to the network helps to improve the direct sound plus early reflection that is hard to predict directly. In the future, we would like to address non-stationary (e.g. moving) speakers.

7. REFERENCES

[1] Y. Zhao, D. L. Wang, B. Xu, and T. Zhang, "Late reverberation suppression using recurrent neural networks with long short-term memory," in *Proc. ICASSP*. IEEE, 2018, pp. 5434–5438.

[2] K. Han, Y. Wang, and D. L. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. ICASSP*. IEEE, 2014, pp. 4628–4632.

[3] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio Speech and Lang. Process.*, vol. 24, no. 3, pp. 483–492, 2016.

[4] D. S. Williamson and D. L. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process. (IEEE TASLP)*, vol. 25, no. 7, pp. 1492–1501, 2017.

[5] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio Speech Lang. Process*, vol. 18, no. 7, pp. 1717–1731, 2010.

[6] T. Yoshioka and T. Nakatani, "Generalization of multichannel linear prediction methods for blind mimo impulse response shortening," *IEEE Trans. Audio Speech Lang. Process*, vol. 20, no. 10, pp. 2707–2720, 2012.

[7] M. Delfarah and D. L. Wang, "Deep learning for talker-dependent reverberant speaker separation: An empirical study," *IEEE/ACM Trans. Audio Speech and Lang. Process.*, vol. 27, no. 11, pp. 1839–1848, 2019.

[8] Y. Sun, W. Wang, J. Chambers, and Syed M. Naqvi, "Two-stage monaural source separation in reverberant room environments using deep neural networks," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process. (IEEE TASLP)*, vol. 27, no. 1, pp. 125–139, 2018.

[9] J. F. Santos and T. H. Falk, "Speech dereverberation with context-aware recurrent neural networks," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process. (IEEE TASLP)*, vol. 26, no. 7, pp. 1236–1246, 2018.

[10] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[11] 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.

[12] L.-H. Kim and M. Hasegawa-Johnson, "Toward overcoming fundamental limitation in frequency-domain blind source separation for reverberant speech mixtures," in *2010 Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers*. IEEE, 2010, pp. 542–545.

[13] T. Mei, J. Xi, F. Yin, A. Mertins, and J. F. Chicharo, "Blind source separation based on time-domain optimization of a frequency-domain independence criterion," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 6, pp. 2075–2085, 2006.

[14] W. Wu and L. Zhang, "A new method of solving permutation problem in blind source separation for convolutive acoustic signals in frequency-domain," in *2007 International Joint Conference on Neural Networks*. IEEE, 2007, pp. 1237–1242.

[15] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 2, pp. 109–116, 2003.

[16] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, et al., "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge," in *Reverb workshop*, 2014.

[17] D. S. Williamson and D. L. Wang, "Speech dereverberation and denoising using complex ratio masks," in *Proc. ICASSP*. IEEE, 2017, pp. 5590–5594.

[18] S Patro and K. K. Sahu, "Normalization: A preprocessing stage," *arXiv preprint arXiv:1503.06462*, 2015.

[19] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process. (IEEE TASLP)*, vol. 22, no. 12, pp. 1849–1858, 2014.

[20] J. S. "Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. " Dahlgren, "Darpa timit acoustic phonetic continuous speech corpus," in *Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html*, 1993.

[21] E. Habets, "Room impulse response generator (http://home.tiscali.nl/ehabets/rir generator.html)," 2010.

[22] ITU-R, "Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," 2001, p. 862.