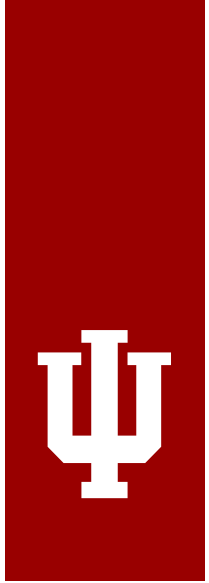


A RETURN TO DEREVERBERATION IN THE FREQUENCY DOMAIN USING A JOINT LEARNING APPROACH

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2020

May 8th, 2020

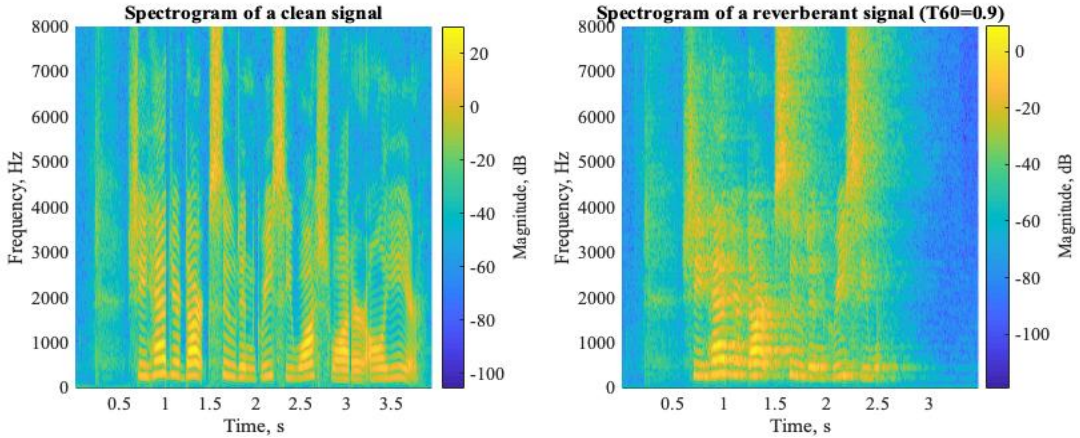
Yuying Li and Donald S. Williamson

LUDDY

SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING

Why Dereverberation?

- Reverberation degrades perceptual speech quality and intelligibility as sound reflections obscure signal structure
- The presence of reverberation may undermine our listening experience
- The reverberation could affect many applications, such as, hearing aids, automatic recognition (ASR) and speaker identification



Previous Studies

- LSTM predicts reflections in time-frequency (T-F) domain. The prediction is subtracted from the reverberant speech signal [5]
  - Pro: LSTM captured time information that could help predicting the late reflections
  - Con: performed on magnitude ignored phase, which could affect the results.

- DNN predicts complex ratio mask (cIRM), where the mask enhances the magnitude and phase [3]
  - Pro: cIRM combined both magnitude and phase information
  - Con: DNN could not capture the continuous time information

- Weighted prediction error (WPE), estimates an inverse filter that is subtracted from the reverberated speech [6, 7]
  - Pro: handles both multi-channel and single-channel situation, more generalization
  - Con: did not perform as well as DNN approaches

- Another approach that uses T-F domain features to predict a dereverberation mask (DM) and IRM [8]
  - Pro: can do source separation and dereverberation at the same time
  - Con: lack of time information

Frequency vs. T-F domain

- 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
- Many previous studies performed in the frequency-domain (e.g. independent component analysis (ICA)) [12]
  - They require assumptions to hold in different environments, which is not often the case.
  - A frequency-domain deep learning-based approach, however, may not need to make these assumptions

Problem formulation

- **Objective**: remove the late reflections from the corresponding reverberant speech signal by operating in the frequency domain

- **Prediction target**: direct plus early RIR in frequency domain

- **Method**: joint-LSTM network to predict both direct plus early RIR and late RIR in frequency domain

Introduction: Reverberation

- A reverberant speech signal can be computed as the convolution of a clean speech signal s(t) with a room impulse response (RIR), h(t):

$$x(t) = s(t) * h(t)$$

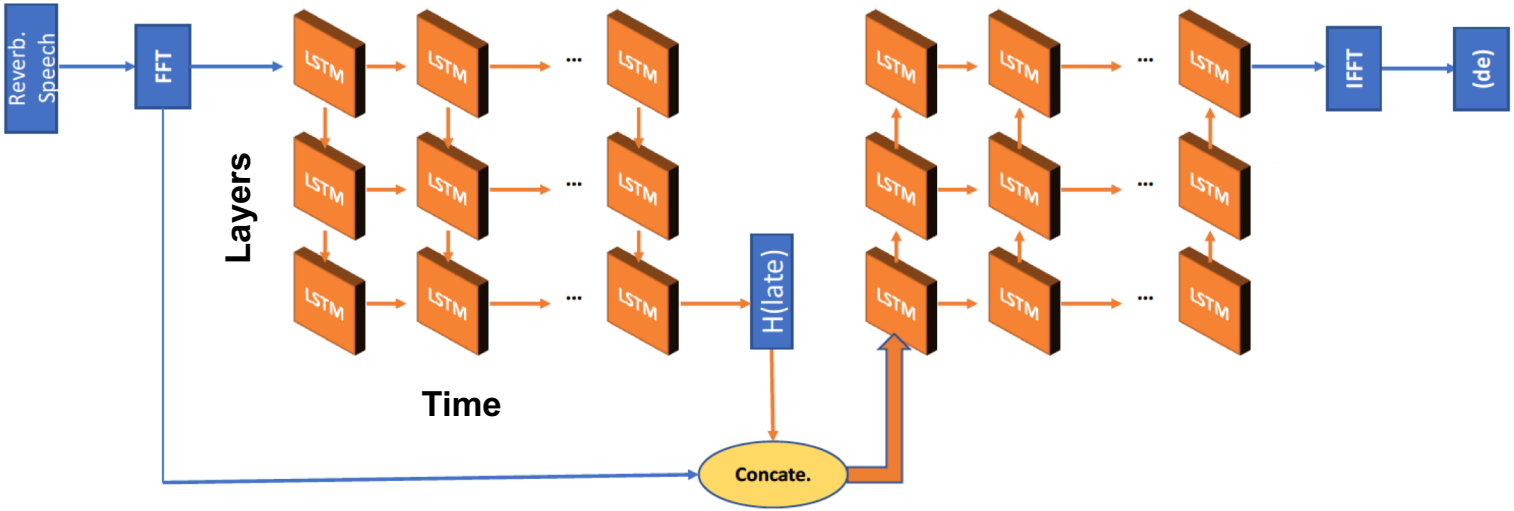
- RIR decomposed into direct, early and late:

$$h(t) = h_d(t) + h_e(t) + h_l(t)$$

- Reverberant signal is defined as direct sound plus early and late reflections:

$$x(t) = s(t) * h_d(t) + s(t) * h_e(t) + s(t) * h_l(t)$$

Network Architecture



Network Architecture

- The first part of the network predicts the late RIR in frequency domain
  - The number of neurons is set to 2048 for each LSTM layer
  - Three fully connected (FCN) layers

- The second part predicts the direct plus early RIR in frequency domain
  - The number of neurons is set to 4096 for first 2 LSTM layers
  - The number of neurons is set to 2048 for the last LSTM layer
  - Three FCN layers

- Rectified linear activation function used throughout
- Adam optimizer
- Learning rate: 0.0001
- Objective function: MSE

Features

- Given a time domain reverberant signal y(t)
  - Compute the 1024-point discrete Fourier transform (DFT)
  - Concatenate the real and imaginary components of the 1024-point DFT as input

$$Y(m) = \begin{bmatrix} Y_i\,(1)\; Y_i\,(2) \ldots Y_i\;(N) \\ Y_r\,(1)\; Y_r\,(2) \ldots Y_r\;(N) \end{bmatrix}$$

$$Y = \{Y(1)Y(2) \ldots Y(N_s)\}$$

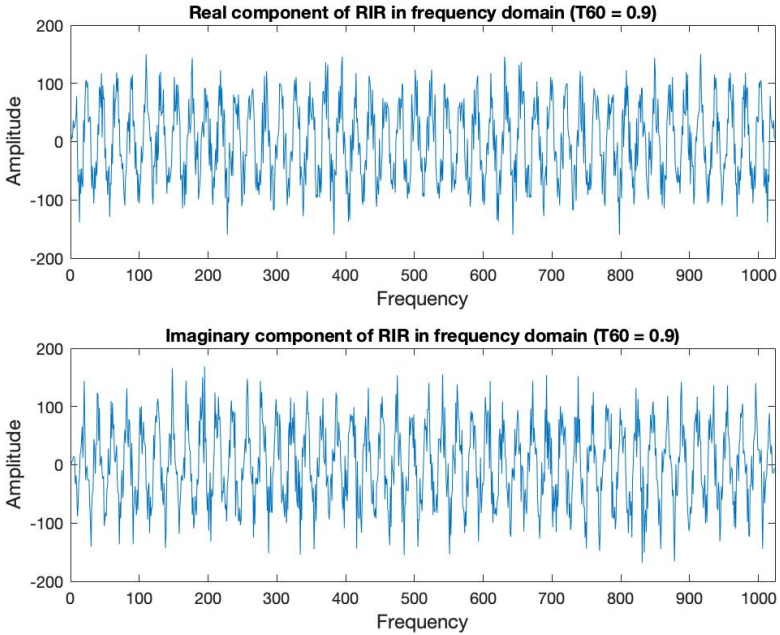
  - Most approaches ignore phase information
  - Recently, research has shown that including phase information improves results [3]
  - Our approach, given the real and imaginary components, means magnitude and phase information are addressed

Training Labels

- Predict transfer functions of the RIRs instead of speech:
  - Transform the direct plus early RIR ($h_{de}(t)$) and the late RIR ($h_l(t)$) into 1024- point DFTs (N = 1024)
  - Concatenate the real and imaginary components of the resulting DFT into one

$$H_l(m) = \{H_l(1)\, H_l(2) \ldots H_l(N)\}$$

$$H_{de}(m) = \{H_{de}(1)\, H_{de}(2) \ldots H_{de}(N)\}$$



Objective Function

- Trained using standard back propagation algorithm with mean-square error cost function

$$\frac{1}{2N}\sum_{f}\left[\left(\widehat{H}_{l}^{r}(f)-H_{l}^{r}(f)\right)^{2}+\left(\widehat{H}_{l}^{i}(f)-H_{l}^{i}(f)\right)^{2}\right]$$
$$+\frac{1}{2N}\sum_{f}\left[\left(\widehat{H}_{de}^{r}(f)-H_{de}^{r}(f)\right)^{2}+\left(\widehat{H}_{de}^{i}(f)-H_{de}^{i}(f)\right)^{2}\right]$$

- N is the input dimensions, which is 1024 units

Experiments

- Dataset: TIMIT corpus
- Randomly select 3000, 1000, and 1000 sentences to construct training, validation and testing datasets
- Simulate RIRs from 5 different rooms via image method [4]. The distance between the receiver and the speaker is set to 1m in all cases
- Select 3 different reverberation times: 0.3s, 0.6s and 0.9s
- 1500 different RIRs for training set, 500 different RIRs for validation set, and another 500 RIRs for testing set

Evaluation

- We convolve the estimated direct plus early RIR with clean speech, we call it estimated speech
- Compare the estimated speech with the true direct plus early speech signal
- Perceptual evaluation of speech quality (PESQ) [9]
  - PESQ score ranges from -0.5 to 4.5
- Short-Time Objective intelligibility (STOI) [10]
  - STOI measure ranges from 0 to 1
- Signal to distortion (SDR) [11]
  - No specific range
- Overall, higher values indicate better performance

Results

- SDR:
  - cIRM performs best at T_{60} of 0.3
  - Joint-LSTM outperforms the comparison approaches as the T_{60} increases
- STOI:
  - Proposed approaches has minor improvements compared with mixture, but not as good as cIRM and IRM
- PESQ:
  - Proposed approaches outperforms the baseline approaches
- Overall, our joint learning approaches perform better as T_{60} increases
- More noticeably, the results reveal that frequency-domain processing often outperforms T-F domain processing.

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 [3]	**7.90**	7.32	7.95	0.73	**0.70**	**0.71**	**2.28**	2.03	2.09
IRM [2]	7.62	7.27	7.54	0.72	**0.70**	0.70	2.23	2.00	2.01
Spectral Mapping [1]	7.28	7.25	7.48	0.71	0.69	0.68	2.01	1.97	1.90

Conclusion

- Our approach managed to extract RIR frequency information and enhance the reverberant signal
- Our approach enhanced reverberant speech in the complex domain by handling both real and imaginary information
- Our approach deviates from recent methods by processing in the frequency domain
- The joint-learning method, by adding predicted late information to the network, helps to improve the direct sound plus early reflection that is hard to predict directly

References

1) K. Han, Y. Wang, and D. L. Wang, “Learning spectral mapping for speech dereverberation,” in Proc. ICASSP. IEEE, 2014, pp. 4628–4632.

2) M. Delfarah and D. L. Wang, “Deep learning for talkerdependent reverberant speaker separation: An empirical study,” IEEE/ACM Trans. Audio Speech and Lang. Process., vol. 27, no. 11, pp. 1839–1848, 2019.

3) D. S.Williamson and D. L.Wang, “Speech dereverberation and denoising using complex ratio masks,” in Proc. ICASSP. IEEE, 2017, pp. 5590–5594.

4) E. Habets, “Room impulse response generator (http://home.tiscali.nl/ehabets/rir generator.html),” 2010.

5) 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.

6) 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.

7) 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.

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) 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.

10) C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time frequency weighted noisy speech,” IEEE Trans. on Audio, Speech, and Lang. Proc., vol. 19, pp. 2125–2136, 2011.

11) E. Vincent, R. Gribonval, and C. Fevotte, “Performance mea- ´ surement in blind audio source separation,” IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 4, Jul. 2006.

12) 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.

Thank You

The ASPIRE Research Group

https://aspire.sice.indiana.edu/