DEREVERBERATION BASED ON BIN-WISE TEMPORAL VARIATIONS OF COMPLEX SPECTROGRAM

Tzu-Hao Chen, Chun Huang, and Tai-Shih Chi

Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu 300, Taiwan

ABSTRACT

Humans analyze sounds not only based on their frequency contents, but also on the temporal variations of the frequency contents. Inspired by auditory perception, we propose a deep neural network (DNN) based dereverberation algorithm in the rate domain, which presents the temporal variations of frequency contents, in this paper. We show convolutional noise in the time domain can be approximated to multiplicative noise in the rate domain. To remove the multiplicative noise, we adopt the rate-domain complex-valued ideal ratio mask (RDcIRM) as the training target of the DNN. Simulation results show that the proposed rate-domain DNN algorithm is more capable of recovering high-intelligible and high-quality speech from reverberant speech than the compared state-of-the-art dereverberation algorithm. Hence, it is highly suitable for speech applications involving human listeners.

***Index Terms*—** Dereverberation, deep neural network, ideal ratio mask, modulation spectrum.

1. INTRODUCTION

In our daily lives, intelligibility and quality of speech is inevitably degraded due to the reverberant effect of surrounding environments. Therefore, it is important to recover clean speech from reverberant speech for many applications, such as automatic speech recognition (ASR), hearing aids, and voice transmission.

The reverberant speech is formed by convolving clean speech with the room impulse response (RIR), whose length depends on the size and the interior furnishings of the room and is usually not short comparing with the length of an utterance. Therefore, to recover clean speech from reverberant speech by deconvolution becomes a difficult task. Several single-channel dereverberation algorithms have been proposed. Some methods are based on the idea of inverse filtering [1][2] and some are based on the idea of spectral suppression [3][4]. Even if the involved RIR is known, the inverse filtering approach does not always guarantee to work due to the non-causality property of the problem. In contrast, reverberant speech comprises of early reflections and late reverberation and the spectral suppression methods focus on diminishing late reverberation. The fact that keeping early reflections somehow improves speech intelligibility [5][6] backs up the approach of spectral suppression.

During recent years, deep learning has been widely used in many research fields and provided a performance boost over conventional methods. For de-noise and dereverberation tasks, a deep neural network (DNN) can be trained as a regression model to convert a noisy magnitude spectrogram back to a clean one. However, cleaning up only the magnitude spectrogram does not recover a high quality speech signal if the phase spectrogram is still damaged. Since the reverberant effect seriously degrades the phase spectrogram, the magnitude and the phase spectrograms have to be cleaned up at the same time for dereverberation methods to recover high quality speech.

Human hearing demonstrates great capability of dealing with noise and reverberation. Therefore, it is intuitive to adopt certain properties of hearing perception for dereverberation. Based on neuro-physiological data recorded on the auditory cortex, an auditory model has been proposed in [7]. In the model, the brain treats the spectrogram of the sound as a 2D image and further analyzes it in the spectro-temporal modulation domain, which is the 2D Fourier transform of the time-frequency domain (where the spectrogram resides). In other words, the brain analyzes the sound in a double Fourier transform fashion (a double wavelet transform fashion, to be precise, due to the constant-Q selectivity of auditory-related neurons). The temporal modulation domain is referred to as the **rate** domain while the spectral modulation domain is referred to as the **scale** domain. Under the framework of the double Fourier transform analysis, reverberant speech resulted from the convolution with a RIR in the time domain can be approximately expressed in the rate domain as the modulation content of clean speech corrupted with a RIR-related multiplicative noise. Therefore, in this paper, we propose a rate-domain based DNN dereverberation algorithm to diminish the RIR-related multiplicative noise in the rate-domain. In addition, to recover high quality speech, we utilize the rate domain complex-valued ideal ratio mask (RDcIRM) as the training target of the DNN to simultaneously deal with magnitude and phase components in the rate and the spectrogram domains. Once the rate domain is cleaned up, the time domain waveform can be obtained in the double inverse Fourier transform fashion.

The rest of this paper is organized as follows. In Section 2, we formulate the problem and show the transformation of the reverberant signal from the time domain to the rate domain. In Section 3, we introduce the RDcIRM and propose the rate-domain DNN for dereverberation. Experiment results are demonstrated in Section 4 and conclusion is given in Section 5.

2. PROBLEM FORMULATION

The single-channel reverberant speech signal $r(n)$ is obtained by convolving the clean signal $x(n)$ with the N-tap RIR $h(n)$ in the time domain as:

$$r(n) = \sum_{p=0}^{N-1} h(p)x(n-p) \tag{1}$$

The tap number N of the RIR is usually quite large in the time domain such that the time-domain deconvolution from $r(n)$ back to $x(n)$ becomes a very difficult task. The spectrogram of the reverberant signal can be obtained by applying the short-time Fourier transform (STFT) to the above equation. After certain approximations, the complex spectrogram of the reverberant signal can be obtained by the bin-wise convolution of the complex spectrogram of the clean signal and the complex spectrogram of the RIR [8], i.e.,

$$R(k,m) = \sum_{p=0}^{N_h-1} H(k,p)X(k,m-p) \tag{2}$$

where $R(k,m)$, $X(k,m)$, and $H(k,m)$ respectively denote the complex spectrograms of the reverberant signal, the clean signal, and the RIR. The k and m denote the frequency bin and the time frame indexes, and N_h is the length of the spectrogram of RIR counted in frames.

Obviously, N_h is much smaller than N. In other words, by these approximations, we transfer the time-domain convolution with a long-tap RIR into a frequency bin-wise convolution with a short-tap RIR in the STFT domain. Furthermore, we create the rate-domain representation by taking another frequency bin-wise Fourier transform on the complex spectrogram as Eq. (3).

$$r_{rate}(k,n') = \mathcal{F}_m\{R(k,m)\} = h_{rate}(k,n')x_{rate}(k,n') \tag{3}$$

where r_{rate}, x_{rate}, and h_{rate} denote the complex-valued rate-domain representations of the reverberant signal, the clean signal, and RIR, respectively. The n' is the rate bin index and the $\mathcal{F}_m\{\cdot\}$ denotes the frequency bin-wise Fourier transform of the complex-valued frame series.

It has been shown that the complex-valued ideal ratio mask (cIRM) of the spectrogram can be used as the training target of a STFT-domain DNN to enhance speech [9]. Simulation results showed that this kind of DNN can diminish additive noise effectively but not reverberant noise. We think the reason is that the time-domain convolutional noise is still a convolutional noise on the spectrogram as per Eq. (2). And multiplying a gain is effective in reducing additive noise (such as the Wiener filter) and multiplicative noise, but not convolutional noise. Therefore, applying a mask on the spectrogram won't do any good in reducing reverberation. Motivated by hearing perception, we transfer the dereverberation task from the STFT domain to the rate domain where we deal with RIR-related multiplicative noise. In the next section, we propose a rate-domain DNN which learns RDcIRM to suppress the multiplicative noise in the rate domain for dereverberation.

3. PROPOSED ALGORITHM

In this section, we describe the proposed rate-domain DNN-based dereverberation algorithm with its key elements, the RDcIRM and the re-synthesis procedure.

3.1. Rate domain complex-valued ideal ratio mask

The RDcIRM is derived from the rate-domain representations of the clean and the reverberant signals. Here, we describe the procedures to produce the rate-domain representation using the clean signal as an example. The 16 kHz sampled clean signal $x(n)$ is first transferred into the spectrogram $X(k,m)$ using a 20-ms window with a 10-ms shift. The 1024-point discrete Fourier transform (DFT) is performed for each frame such that the spectrogram has 513 frequency bins (including the bin of DC). After obtaining the spectrogram $X(k,m)$, we apply a 512-point DFT to each frequency bin as Eq. (3) to get the rate-domain representation $x_{rate}(k,n') \in \mathcal{C}^{513*512}$. Since the complex-valued spectrogram $X(k,m)$ is considered, all the 512 points in the rate domain have to be preserved for further processing. The RDcIRM IM_{rate}, which is the training target of the DNN, is then obtained by the element-wise division between the rate-domain representations of the clean and the reverberant signals as follows

$$IM_{rate}(k,n') = \frac{x_{rate}(k,n')}{r_{rate}(k,n')} \tag{4}$$

Similar to the approach of training a DNN to learn the cIRM from magnitude spectrograms in a frame-by-frame fashion [9], we trained a DNN to learn the RDcIRM from rate-domain magnitude representations in a rate-by-rate fashion. Fig. 1 shows the structure of the proposed DNN to learn the RDcIRM. For each rate, the magnitude representations across 513 frequency bins were used as input features to the DNN, with three 1024-neuron hidden layers, while the training target consisted of the real part and the imaginary part of the RDcIRM.

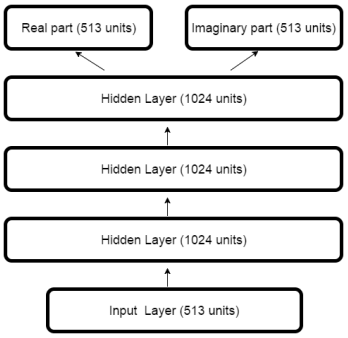


Fig. 1. Structure of the proposed DNN for RDcIRM training.

3.2. Re-synthesis

For dereverberation, we first obtained the estimated RDcIRM $I\hat{M}_{rate}(k,n')$ from the trained DNN and then conducted the element-wise multiplication to the reverberant signal in the rate domain as in Eq. (5).

$$\hat{x}_{rate}(k,n') = r_{rate}(k,n') \cdot I\hat{M}_{rate}(k,n') \tag{5}$$

where $\hat{x}_{rate}$ denotes the estimated clean signal in the rate domain. In addition, the magnitude of $\hat{x}_{rate}$ was thresholded as in Eq. (6):

$$|\hat{x}_{rate}(k,n')| = \begin{cases} \overline{|\hat{x}_{rate}(k,n')|} & if\ |\hat{x}_{rate}(k,n')| \geq \gamma, \\ |\hat{x}_{rate}(k,n')| & otherwise. \end{cases} \tag{6}$$

where $\overline{|\hat{x}_{rate}(k,n')|}$ denotes the averaged magnitude of the estimated rate representation of the clean signal, and γ denotes the threshold which was empirically set to 100 in this study.

Once the estimated rate-domain representation of the clean signal $\hat{x}_{rate}$ was obtained, the frequency bin-wise inverse Fourier transform was performed to transfer it back to the STFT domain. Then, the estimated spectrogram was converted to a time-domain waveform by the inverse STFT with the overlap-and-add method. Fig. 2 shows the magnitude spectrograms of the original clean signal, the reverberant signal, and the reconstructed clean signal. The central panel clearly shows the smeared time-frequency structures of speech due to reverberation. The bottom panel shows the smeared structures have been successfully restored to their original shapes but with some additive noise probably due to clipping in the rate domain as in Eq. 6. This artificial additive noise can be easily removed by conventional methods such as the Wiener filter [10]. Therefore, we adopted the Wiener filter as a post-processing method to our dereverberation algorithm in our experiments.

4. EXPERIMENT RESULTS

Our main goal is to develop a dereverberation algorithm for human-listening applications, not for machine-listening applications. Therefore, in addition to the frequency-weighted signal-to-reverberation ratio (SRR$_{fw}$) [11], we adopt two other objective measures, the short time objective intelligibility (STOI) [12] and the perceptual evaluation of speech quality (PESQ) [13], to evaluate speech intelligibility and speech quality of re-synthesized speech of the proposed algorithm. The SRR$_{fw}$ computes the averaged signal-to-reverberation ratio over critical bands with different weights as follows:

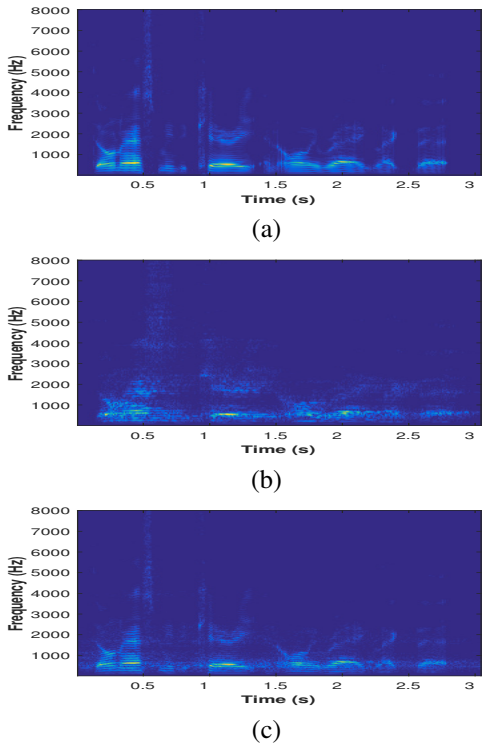


Fig. 2. Dereverberation results. (a) Magnitude spectrogram of original clean speech. (b) Magnitude spectrogram of reverberant speech with T_{60} = 0.9 sec. (c) Magnitude spectrogram of re-synthesized speech after dereverberation.

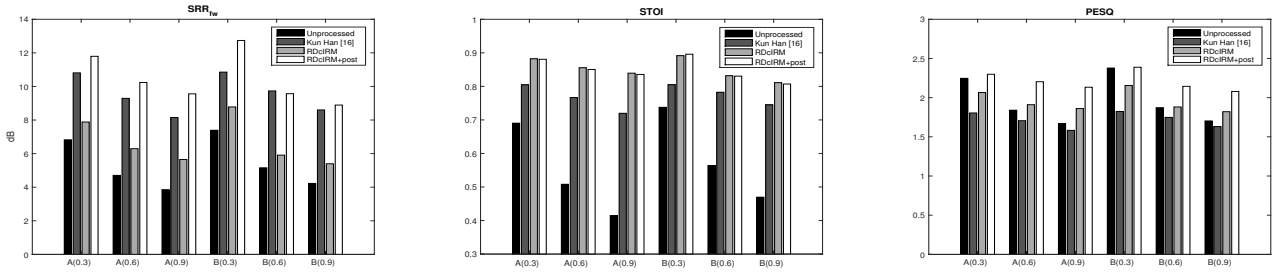
$$SRR_{fw} = \frac{10}{M}\sum_{m=1}^{M}\frac{\sum_{k=1}^{K} w(k,m) log_{10}\frac{|X(k,m)|^2}{(|X(k,m)|-|\hat{X}(k,m)|)^2}}{\sum_{k=1}^{K} w(k,m)} \tag{7}$$

where $|X(k,m)|$ and $|\hat{X}(k,m)|$ denote the magnitude spectrograms of the clean signal and the resynthesized signal, respectively, and K is the total number of frequency bin and M is the total number of frame. The weights $w(k,m)$ can be selected as:

$$w(k,m) = |X(k,m)|^p \tag{8}$$

where p was set to 2 in our experiments to account for the power of the magnitude spectrogram.

In our experiments, the RIR generator [14] was used to generate simulated RIRs, which were convolved with clean utterances to produce reverberant utterances. We generated



(a) SRR_{fw} (b) STOI (c) PESQ

Fig. 3. Performance of dereverberation in terms of three measures. "Unprocessed" denotes original reverberant signals, "Kun Han [16]" denotes the compared state-of-the-art dereverberation system, "RDcIRM" denotes the proposed algorithm without the post processing (Wiener filter), and "RDcIRM + post" denotes the proposed algorithm with the post processing. 6 test conditions from combinations of two rooms (A and B) and three T_{60} (0.3, 0.6 and 0.9 sec) are expressed as Room(T_{60}).

six RIRs from two different rooms (room A and room B) with three different T_{60}, 0.3, 0.6 and 0.9 sec. 200 clean utterances were extracted from the TIMIT corpus [15] and convolved with the six RIRs to constitute the training set of 1200 reverberant utterances. The other different 100 utterances from the TIMIT corpus were used to convolve with the six RIRs to constitute the test set of 600 reverberant utterances.

Fig. 3 shows the average performance of the proposed algorithm and the compared state-of-the-art system, which adopts a DNN to map multiple adjacent frames of the reverberant magnitude spectrogram to the present frame of the clean magnitude spectrogram [16], in terms of three measures under 6 test conditions. Fig. 3(a) shows the proposed RDcIRM method has lower SRR_{fw} scores than the compared system due to the artificial additive noise introduced by the rate-domain clipping. However, the scores of the proposed method can be greatly increased by removing the artificial noise using a simple post-processing Wiener filter. On the other hand, Fig. 3(b) and (c) show the proposed method provides a great advantage over the compared method in terms of STOI and PESQ scores. In addition, these results clearly show that the artificial additive noise of our method only degrades PESQ scores but has no negative impacts on STOI scores. Once the artificial noise being removed by the Wiener filter, the PESQ scores will advance higher as shown in Fig. 3(c). Based on these results, we can conclude that the proposed method is better at preserving speech intelligibility and speech quality so that it is more suitable for human-listening applications. Sound examples from the proposed dereverberation method are available at http://perception.cm.nctu.edu.tw/sound-demo/.

To evaluate the generality of the proposed method, we trained the RDcIRM from conditions of $\text{T}_{60} \in \{0.3, 0.6, 0.9\}$ sec and tested the method for different T_{60}. The STOI scores at all T_{60} conditions shown in Fig. (4) demonstrate the proposed method can fairly preserve speech intelligibility even in unseen T_{60} conditions.

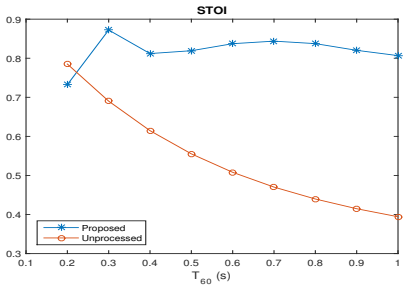


Fig. 4. Generalization test for different T_{60}. The method for the "Proposed" curve is the proposed RDcIRM method.

5. CONCLUSION

In this paper, we propose a rate-domain dereverberation algorithm. Inspired by auditory perception, the complex spectrogram of the speech signal is first bin-wise transferred to its temporal-variation domain, the rate domain. In this way, convolutional noise in the time domain can be approximated as multiplicative noise in the rate domain. We then build the RDcIRM as the training target of a DNN to remove the multiplicative noise for dereverberation. Compared with a state-of-the-art dereverberation system [16], the proposed algorithm can produce de-reverberated speech with higher speech intelligibility and speech quality scores such that it is more suitable for human-listening applications.

6. ACKNOWLEDGEMENT

This research is supported by the Ministry of Science and Technology, Taiwan under Grant No MOST 105-2221-E-009-152-MY2.

7. REFERENCES

[1] D. T. Fee, C. F. N. Cowan, and S. Bilbao, "Predictive deconvolution and kurtosis maximization for speech dereverberation," in *Signal Processing Conference, 2006 14th European*. IEEE, 2006, pp. 1–5.

[2] M. J. Daly and J. R. Reilly, "Blind deconvolution using bayesian methods with application to the dereverberation of speech," in *Proc. ICASSP*. IEEE, 2004, vol. 2, pp. 1009–1012.

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

[4] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation," in *Proc. ICASSP*. IEEE, 2006, vol. 1, pp. 817–820.

[5] N. Roman and J. Woodruff, "Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold," *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1707–1717, 2013.

[6] Y. Hu and K. Kokkinakis, "Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. EL22–EL28, 2014.

[7] T. Chi, P. Ru, and S. A. Shamma, "Multi-resolution spectro-temporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[8] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *Trans. Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

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

[10] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. ICASSP*. IEEE, 1996, vol. 2, pp. 629–632.

[11] 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, no. 5, pp. 3387–3405, 2009.

[12] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *Trans. Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[13] A. W. Rix, J. G. Beerends, and M. P. Hollier, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, vol. 2, pp. 749–752.

[14] E. A. P. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, pp. 1, 2006.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, and J. G. Fiscus, "Timit acoustic-phonetic continuous speech corpus," *Linguistic data consortium, Philadelphia*, vol. 33, 1993.

[16] K. Han, Y. Wang, D. L. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *Trans. Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.