Applied Acoustics 134 (2018) 154–159

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Guided spectrogram filtering for speech dereverberation

Chengshi Zheng[a], Zheng-Hua Tan[b], Renhua Peng[a,*], Xiaodong Li[a]

[a] Key Laboratory of Sound and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, 100190 Beijing, China
[b] Department of Electronic Systems, Signal and Information Processing Section, Aalborg Univeristy, Aalborg 9220, Denmark

ARTICLE INFO

Keywords:
Guided filter
Guided image filtering
Spectrogram
Dereverberation

ABSTRACT

Guided filtering is a computationally efficient and powerful technique used in image processing applications, such as edge-preserving smoothing, details enhancing and single image dehazing. In this paper, we propose a novel single channel speech dereverberation method using guided spectrogram filtering by considering a speech spectrogram as an image. The proposed method requires neither room acoustic parameter estimation nor late reverberant spectral variance estimation. Objective test results show the validity of the guided spectrogram filtering method for speech dereverberation. Compared with state-of-the-art speech dereverberation methods, the proposed method has better performance in terms of perceptual evaluation of speech quality (PESQ), speech-to-reverberation modulation energy ratio (SRMR) and short-time objective intelligibility (STOI) in most cases.

1. Introduction

In reverberant environments, speech quality and speech intelligibility may degrade dramatically due to acoustic reverberation. Also, speech recognition often fails in highly reverberant conditions. Speech dereverberation is important for hands-free speech communication systems and human-machine speech interfaces [1–4]. Numerous effective methods have already been proposed to reduce late reverberation components in the last half century [5–17].

Conventional single-channel speech dereverberation methods often need to estimate the late reverberant spectral variance [7–17]. For this purpose, some room acoustic parameters usually need to be estimated blindly, such as the reverberation time (T_{60}) or the damping constant. Note that some methods can estimate the late reverberant spectral variance without estimating any room acoustic parameters. In [11], the late reverberant spectral variance is estimated by using long-term multi-step linear prediction. In [12], however, it is estimated by exploiting long-term correlation.

It is well-known that acoustic reverberation has impact on clean speech spectrograms. If a clean speech spectrogram is considered as a clean image, its corresponding reverberant speech spectrogram can be considered as a corrupted version of the clean image. Single image denoising is a hot topic in image processing and numerous algorithms have been proposed in recent years, such as the bilateral filter and the guided filter [18–21]. Since spectrograms contain very useful information of the represented signals, they are used for various purposes, such as biological signals denoising [22,23], speech enhancement [24,25], speech recognition [26], fundamental frequency extraction [27], sound classification and speaker identification [28–30].

In this paper, a guided spectrogram filtering method is proposed to reduce acoustic reverberation for the following three considerations. First, as pointed out in [21], the guided filter has shown its effectiveness and efficiency in many computer vision and computer graphics applications. Second, it does not have unwanted *gradient reversal* artifacts near edges that the bilateral filter may have. Third, it has better performance and much less computational cost than the lateral filter [18].

The remainder of this paper is organized as follows. Section 2 formulates the problem. Section 3 presents the proposed guided spectrogram filtering method and the detailed description of reconstructing the time-domain enhanced speech. Experimental results and conclusions are given in Section 4 and Section 5, respectively.

2. Problem formulation

In reverberant environments, a microphone signal can be modeled as

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

where $s(n)$ is discrete-time domain clean speech and $h(n,m)$ is the linear transfer function from the source signal $s(n)$ to the microphone. $h(n,m)$ could be time-varying or time-invariant. For a linear time-invariant system, $h(n,m) = h(n-m) = h(\tau)$ holds, where $\tau = n-m$. When the geometry of the talker and the microphone does not change rapidly, it can be assumed that the transfer function from $s(n)$ to the

* Corresponding author.
E-mail address: pengrenhua@mail.ioa.ac.cn (R. Peng).

https://doi.org/10.1016/j.apacoust.2017.11.016
Received 1 September 2017; Received in revised form 20 October 2017; Accepted 16 November 2017
Available online 02 February 2018
0003-682X/

microphone is approximately linear time-invariant. Most of blind system identification-based algorithms assume that the transfer function is linear time-invariant, or approximately linear time-invariant, and thus their performances may degrade a lot for a moving talker [4].

Bradley et al. have shown that early reflections that reach the microphone in approximately the first 50–100 ms after the direct path are beneficial to overall speech intelligibility [31]. Hu et al. further show that early reflections neither improve nor decrease overall speech perception [32]. However, it is well-known that late reflections may reduce both speech quality and speech intelligibility, so that they need to be suppressed. For this reason, (1) is rewritten as

$$x(n) = \sum_{m=n-D_h+1}^{n} h(n,m)s(m) + \sum_{m=-\infty}^{n-D_h} h(n,m)s(m), \tag{2}$$

where D_h denotes the filter length of early reflections. (2) is further written as

$$x(n) = z_{\mathscr{E}}(n) + z_{\mathscr{L}}(n), \tag{3}$$

where $z_{\mathscr{E}}(n) = \sum_{m=n-D_h+1}^{n} h(n,m)s(m)$ denotes the early speech component and $z_{\mathscr{L}}(n) = \sum_{m=-\infty}^{n-D_h} h(n,m)s(m)$ denotes the late reverberant speech component. In this paper, only the late reverberation needs to be suppressed, while the early speech component is unaltered.

Applying the short-time Fourier transform (STFT) to (3), we get

$$X(k,l) = Z_{\mathscr{E}}(k,l) + Z_{\mathscr{L}}(k,l), \tag{4}$$

where $X(k,l) = \sum_{n=0}^{K-1} x(n + lR)w(n)e^{-j\frac{2\pi}{K}nk}$ is the STFT of $x(n)$. k and l are, respectively, the frequency bin index and the frame index. K is the frame length and R is the frame shift. Analogous to $X(k,l)$,$Z_{\mathscr{E}}(k,l)$ and $Z_{\mathscr{L}}(k,l)$ correspond to the STFTs of $z_{\mathscr{E}}(n)$ and $z_{\mathscr{L}}(n)$, respectively. Spectrograms of the microphone signal, the early speech component, and the late speech component are denoted as $\mathscr{X}(k,l) = 20\log_{10}(|X(k,l)|^2)$,$\mathscr{Z}_{\mathscr{E}}(k,l) = 20\log_{10}(|Z_{\mathscr{E}}(k,l)|^2)$ and $\mathscr{Z}_{\mathscr{L}}(k,l) = 20\log_{10}(|Z_{\mathscr{L}}(k,l)|^2)$, respectively. Each spectrogram can be regarded as an image, where each time-frequency bin is a pixel.

Before giving an intuitive interpretation, we propose to normalize each spectrogram first. Taking $\mathscr{X}(k,l)$ as an example, we have

$$\widetilde{\mathscr{X}}(k,l) = \frac{\overline{\mathscr{X}}(k,l)}{\max\{\overline{X}\}} \tag{5}$$

where $\max\{\overline{X}\}$ extracts the maximum value of the matrix $\overline{X}$. The element of $\overline{X}$ is given by

$$\overline{\mathscr{X}}(k,l) = \widehat{\mathscr{X}}(k,l) - \min\{\widehat{X}\} \tag{6}$$

where $\min\{\widehat{X}\}$ extracts the minimum value of the matrix $\widehat{X}$. The element of $\widehat{X}$ is given by

$$\widehat{\mathscr{X}}(k,l) = \begin{cases} \mathscr{X}(k,l), & \text{if } \mathscr{X}(k,l) \geqslant \mathscr{X}_{\min} \\ \mathscr{X}_{\min}, & \text{otherwise} \end{cases} \tag{7}$$

where $\mathscr{X}_{\min} = \max\{\mathcal{X}\} - 255$.

Fig. 1 shows the impact of the late reverberant component on the normalized spectrogram of a clean speech signal. We use the same clean speech signal as in [17], taken from the TIMIT database [33]. The transfer function between the talker and the microphone is generated by the image method [34], where the length, width and height of the simulated rectangular room are 5 m, 4 m and 3 m, respectively. According to Sabine's reverberation model, the value of the reflection coefficient can be calculated directly when the reverberation time T_{60} is given. As can been seen from Fig. 1, the late reverberant component has significant impact on the normalized spectrogram of the clean speech signal. If the normalized spectrogram of the clean speech signal is regarded as a clean image, the normalized spectrogram of the reverberant and the noisy speech can be seen as this clean image with haze. In this paper, an efficient and effective haze removal algorithm is adopted to extract the early speech component $z_{\mathscr{E}}(n)$ from the microphone signal $x(n)$.

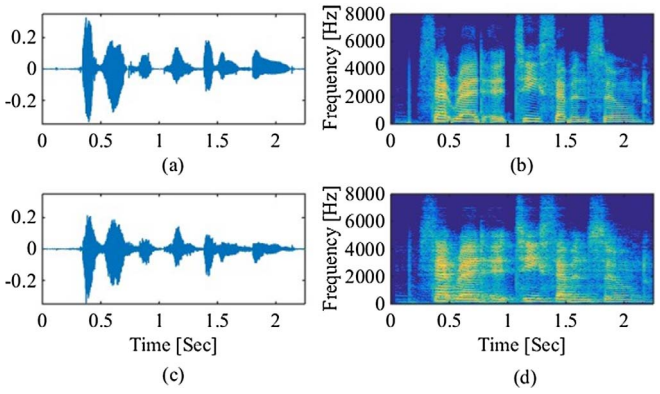

Fig. 1. Waveforms and normalized spectrograms of the clean speech (a), (b), the reverberant speech with T_{60} = 400 ms (c), (d).

3. Guided spectrogram filtering

To the best of our knowledge, the guided image filtering method [21] has not been introduced for speech dereverberation yet. Based on the guided image filtering method, we propose a guided spectrogram filtering method for speech dereverberation. In practice, only the normalized spectrogram of $x(n)$ is available, and thus we only have the guidance spectrogram $\widetilde{\mathcal{X}}$. The element of the filtering input spectrogram $\widetilde{\mathcal{P}}$ can be computed from $\widetilde{\mathcal{X}}$ directly, which is given by

$$\widetilde{\mathcal{P}}(k,l) = \beta\widetilde{\mathcal{P}}(k,l-1) + (1-\beta)\widetilde{\mathcal{X}}(k,l), \tag{8}$$

where β is the forgetting factor and its typical value ranges from 0 to 1. According to [21], the local linear model between $\widetilde{\mathcal{X}}$ and the filtering output spectrogram $\widetilde{\mathcal{Q}}$ is the most important assumption of the guided filter, which leads to

$$\widetilde{\mathcal{Q}} = \mathbf{A}\circ\widetilde{\mathcal{X}} + \mathbf{B}, \tag{9}$$

where $\circ$ denotes the Hadamard product, and both $\mathbf{A}$ and $\mathbf{B}$ are calculated from the guidance spectrogram $\widetilde{\mathcal{X}}$ and the filtering input spectrogram $\widetilde{\mathcal{P}}$, which is given in details herein.

1. Compute the local variance of the guidance spectrogram $\widetilde{\mathcal{X}}$ at each frequency bin k and l, given by

$$\sigma^2_{\widetilde{\mathcal{X}}}(k,l) = m^2_{\widetilde{\mathcal{X}}}(k,l) - (m_{\widetilde{\mathcal{X}}}(k,l))^2, \tag{10}$$

where

$$m^2_{\widetilde{\mathcal{X}}}(k,l) = \sum_{l-r_2}^{l+r_2}\sum_{k-r_1}^{k+r_1} (\widetilde{\mathcal{X}}(k,l))^2/N, \tag{11}$$

and

$$m_{\widetilde{\mathcal{X}}}(k,l) = \sum_{l-r_2}^{l+r_2}\sum_{k-r_1}^{k+r_1} \widetilde{\mathcal{X}}(k,l)/N, \tag{12}$$

where $N = (2r_1 + 1) \times (2r_2 + 1)$ denotes the number of time-frequency bins in computing the local mean and the local mean of square of $\widetilde{\mathcal{X}}$. The parameter r_1 should not be a large value due to that statistical independence assumption over frequency is valid even in reverberant conditions. However, r_2 could be much larger than r_1 when considering that the late reverberant components are smoothed and shifted version of the power spectral densities of the clean speech [8].

2. Compute the local covariance of the guidance spectrogram $\widetilde{\mathcal{X}}$ and the filtering input spectrogram $\widetilde{\mathcal{P}}$ at each frequency bin k and l, given by

$$\mathrm{cov}_{\widetilde{\mathcal{X}}\widetilde{\mathcal{P}}}(k,l) = m_{\widetilde{\mathcal{X}}\widetilde{\mathcal{P}}}(k,l) - m_{\widetilde{\mathcal{X}}}(k,l)m_{\widetilde{\mathcal{P}}}(k,l), \tag{13}$$

where

$$m_{\widetilde{X}\widetilde{\mathcal{P}}}(k,l)=\sum_{l-r_2}^{l+r_2}\sum_{k-r_1}^{k+r_1}\widetilde{X}(k,l)\widetilde{\mathcal{P}}(k,l)/N, \tag{14}$$

and

$$m_{\widetilde{\mathcal{P}}}(k,l)=\sum_{l-r_2}^{l+r_2}\sum_{k-r_1}^{k+r_1}\widetilde{\mathcal{P}}(k,l)/N. \tag{15}$$

3. Calculate $\widetilde{A}(k,l)$ and $\widetilde{B}(k,l)$ using the local covariance, the local variance and the local mean, given by

$$\widetilde{A}(k,l)=\mathrm{cov}_{\widetilde{X}\widetilde{\mathcal{P}}}(k,l)/(\sigma^2_{\widetilde{X}}(k,l)+\varepsilon) \tag{16}$$

and

$$\widetilde{B}(k,l)=m_{\widetilde{\mathcal{P}}}(k,l)-\widetilde{A}(k,l)m_{\widetilde{X}}(k,l) \tag{17}$$

4. Locally smooth $\widetilde{A}(k,l)$ and $\widetilde{B}(k,l)$ as follows

$$A(k,l)=\sum_{l-r_2}^{l+r_2}\sum_{k-r_1}^{k+r_1}\widetilde{A}(k,l)/N, \tag{18}$$

and

$$B(k,l)=\sum_{l-r_2}^{l+r_2}\sum_{k-r_1}^{k+r_1}\widetilde{B}(k,l)/N, \tag{19}$$

where $A(k,l)$ and $B(k,l)$ are, respectively, the element of **A** and that of **B** in (9).

The enhanced spectrogram using the proposed guided spectrogram filtering is given by

$$\mathcal{Y}(k,l)=\max\{\widetilde{X}(k,l)-\alpha\widetilde{Q}(k,l),0\}, \tag{20}$$

where α is a constant value ranging from 0 to $\alpha_{\max}$ and $\widetilde{Q}(k,l)$ is the element of the matrix $\widetilde{Q}$. The same as spectral subtraction-type methods, α can be regarded as a subtraction factor. When $\alpha=0,\mathcal{Y}(k,l)=\widetilde{X}(k,l)$ holds so that all the pixels are unchanged. When α is large enough, $\mathcal{Y}(k,l)\equiv 0$ holds so that all the pixels becomes zero. Therefore, $\alpha_{\max}$ should not be too large. A scaling factor is introduced to normalize the enhanced spectrogram, given by

$$\widetilde{Y}(k,l)=\frac{1}{\max\{\mathcal{Y}\}}\mathcal{Y}(k,l). \tag{21}$$

The normalized enhanced spectrogram is applied to compute the gain function directly, which is given by

$$G(k,l)=\min\left\{\max\left\{\frac{\widetilde{Y}(k,l)}{\widetilde{X}(k,l)},G_{\min}\right\},G_{\max}\right\}, \tag{22}$$

where $G_{\min}\in[-30-10]$ dB is the minimum value of the gain that constrains the residual noise floor and $G_{\max}$ is one.

After obtaining the gain function, an inverse FFT (IFFT) is applied to synthesize the time-domain enhanced speech, which is given by

$$\widetilde{z}_{\mathscr{E}}(n)=\mathrm{IFFT}\{G(k,l)X(k,l)\}, \tag{23}$$

where the overlap-add method is necessary to reconstruct the time-domain enhanced speech generally.

The detailed implementation of the proposed guided spectrogram filtering method is summarized in Algorithm 1. It is worth noting that the proposed method does not need to estimate the late reverberant spectral variance for the purpose of suppressing the late reverberant speech component. Therefore, we need neither the late reverberant spectral variance estimation nor room acoustic parameter estimation to implement the proposed method. Moreover, the implementation of the proposed method is very efficient. This is because the guided filter in Algorithm 1 has an $O(1)$ time implementation for each time and frequency bin [21]. Note that $r_1=r_2$ is chosen in the original guided filter, where only one parameter r is used for image processing in [21]. While considering for speech applications, r_1 in the frequency dimension should be much smaller than r_2 in the time dimension for dereverberation.

Algorithm 1. Guided spectrogram filtering

(1) **Normalized Reverberant Speech Spectrogram Generation**
Input: $x(n)$
Output: $X(k,l)$ and $\widetilde{X}(k,l)$
a) Apply the STFT to obtain $X(k,l)$
b) Use (5)–(7) to calculate $\widetilde{X}(k,l)$
(2) **Guided Filter**
Input: the filtering input spectrogram $\widetilde{X}$,r_1,r_2 and ε
Output: the filtering output spectrogram $\widetilde{Q}$
(a) Compute $\widetilde{\mathcal{P}}$ using (8)
(b) Compute **A** and **B** using (10)–(19)
(c) Compute $\widetilde{Q}$ using (9)
(3) **Speech Reconstruction**
Input: $\widetilde{X}(k,l)$,$\widetilde{Q}(k,l)$ and $X(k,l)$
Output: the time-domain enhanced speech $\widetilde{z}_{\mathscr{E}}(n)$
(a) Compute the enhanced spectrogram $\mathcal{Y}(k,l)$ using (20)
(b) Normalize the enhanced spectrogram using (21)
(c) Calculate the gain function using (22) for dereverberation
(d) Reconstruct $\widetilde{z}_{\mathscr{E}}(n)$ by using IFFT in (23) and the overlap-add method

To show the capability of the proposed guided spectrogram filtering method in an intuitive way, we depict the normalized enhanced speech spectrograms in Fig. 2, where $r_1=1,r_2=2,\alpha=1.0,G_{\min}=-17$ dB, $K=512,R=128$ and $\varepsilon=0.8^2$ are chosen for the sampling rate $f_s=16{,}000$ Hz. Comparing Fig. 2 with Fig. 1 shows that the degraded spectrogram has been enhanced by the proposed method, where the smearing effect of reverberation has already been partially removed.

4. Performance evaluation

To evaluate the performance of the proposed guided spectrogram filtering method, we compare it with the spectral subtraction method presented in [7,35], the GSVD-based single-channel speech dereverberation [6] and the constrained minimum mean square estimation linear prediction-based dereverberation using GSVD [17], where SS,

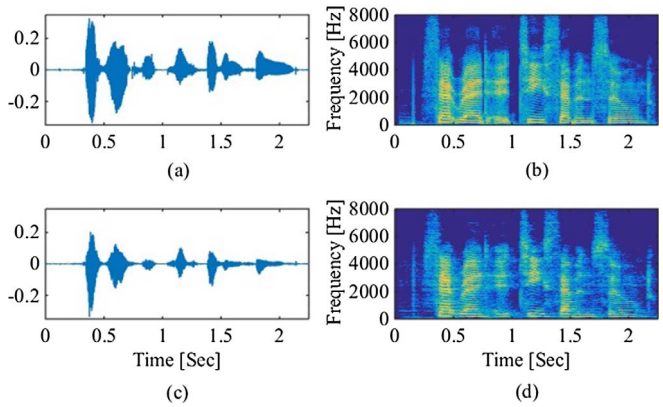

Fig. 2. Waveforms and normalized spectrograms of the clean speech (a), (b), the reverberant speech with $T_{60}=400$ ms enhanced by the proposed guided spectrogram filtering method (c), (d).

GSVD and CMMSE-GSVD are respectively abbreviations of these three existing methods and the proposed method is referred as GSF. To achieve the best performances of these three competing methods, the same as [35], we use the knowledge of the true reverberation times to estimate the late reverberation spectral variance that is need to suppress the reverberant components for these methods, where in practice the reverberation times have to be estimated from the microphone signals blindly.

The clean speech samples are taken from the TIMIT database [33] and sets of measured RIRs for five rooms are used:

1. Room A: 6 m × 6 m × 3 m, fully closed curtains on walls, $T_{60} \approx 0.2$ s
2. Room B: 6 m × 6 m × 3 m, partially closed curtains on walls, $T_{60} \approx 0.4$s
3. Room C: meeting room, 5 m × 3.5 m × 3 m, $T_{60} \approx 0.6$ s
4. Room D: variable reverberation room, 4.5 m × 3.5 m × 3 m, $T_{60} \approx 0.8$ s
5. Room E: lecture hall, 7 m × 11 m × 3 m, $T_{60} \approx 1.0$ s

The detailed description of these sets of measured RIRs can be found in [35–37], where the reverberation times of these five rooms range from 200 ms to 1000 ms. The reverberant signals are generated by convolving the clean speech samples from the TIMIT database with the sets of measured RIRs. The speech to reverberation modulation energy ratio (SRMR) [38,36], the perceptual evaluation of speech quality (PESQ) [39,37] and the short-time objective intelligibility (STOI) [40,41] are chosen to give quantitative comparison results of the proposed method with the three competing methods. The following subsections present the quantitative results and discuss on these results. In all the following results, $r_1 = 1, K = 512, R = 128, \varepsilon = 0.8^2$, and $G_{\min} = -17$ dB are fixed for the sampling rate $f_s = 16{,}000$ Hz, while both r_2 and α are variables, where both of them are set to different values, i.e. $r_2 \in \{2,4,8,16\}$ and $\alpha \in \{0.9,1.0,1.1,1.2\}$.

4.1. PESQ scores

The PESQ scores are shown in Fig. 3. It is obvious that the performance of GSF is highly correlated with r_2,α and the reverberation time T_{60}. When the reverberation time T_{60} is less than 600 ms, higher PESQ scores can be achieved by using a smaller value of r_2, e.g., $r_2 = 2$, while it is better to use a larger value of r_2 when the reverberation time T_{60} is larger than 600 ms, e.g., $r_2 = 4$ or $r_2 = 8$. The reason is that the sound decay rate decreases when the reverberation time T_{60} increases, so more successive frames are necessary to model the late reverberant components. For the subtraction factor α in (20), it suggests that over subtraction is much better than under subtraction, where this conclusion is consistent with the conventional SS methods [1,7]. However, α cannot be too large. PESQ scores significantly reduce when α is larger than 1.2. When comparing the four dereverberation methods, one can find that the proposed GSF method has the highest PESQ scores for $T_{60} \leqslant 800$ ms when $r_2 = 2$ and $\alpha = 1.1$ are chosen. For $T_{60} = 1000$ ms, the CMMSE-GSVD gets the best performance among the four methods, while the proposed method is much better than SS and GSVD.

4.2. SRMR scores

For the SRMR scores, the comparison results are presented in Fig. 4. Comparison on the SRMR scores obtained here is consistent with that

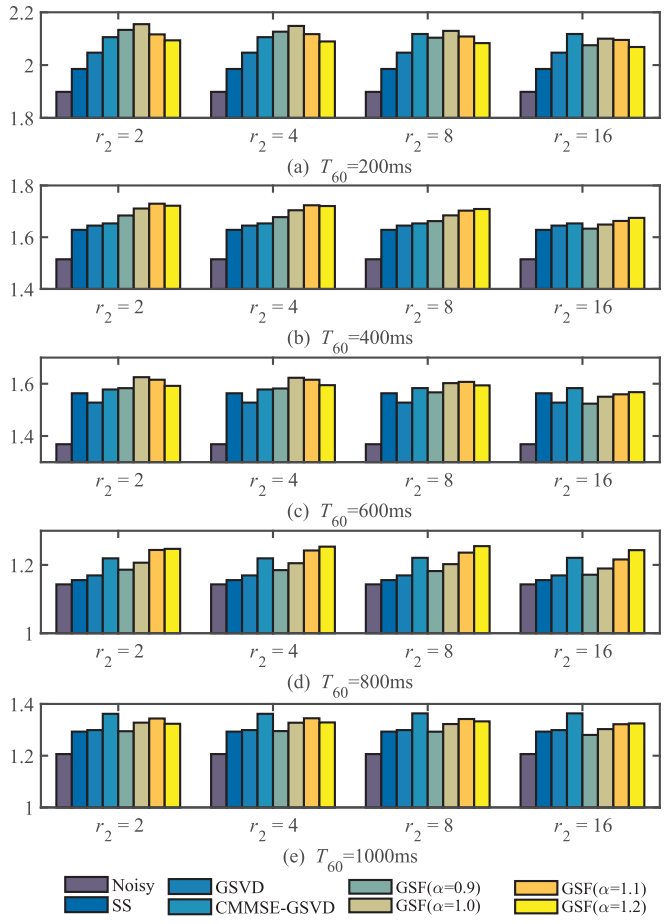

Fig. 3. PESQ scores for Rooms from A to E with $T_{60} = 200$ ms (a), $T_{60} = 400$ ms (b), $T_{60} = 600$ ms (c), $T_{60} = 800$ ms (d), and $T_{60} = 1000$ ms (e), respectively.

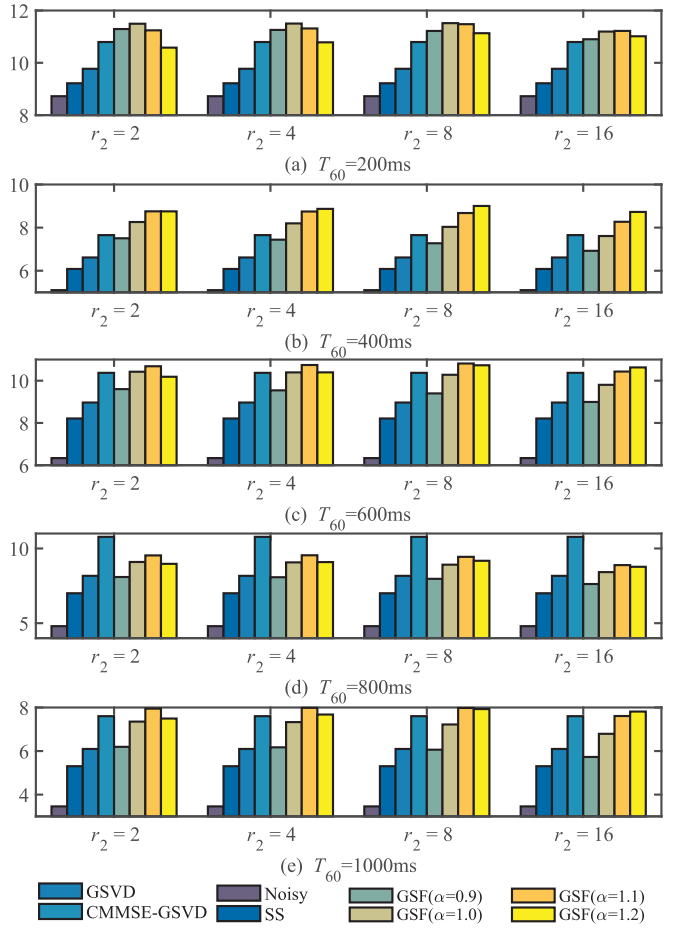

Fig. 4. SRMR scores for Rooms A to E with $T_{60} = 200$ ms (a), $T_{60} = 400$ ms (b), $T_{60} = 600$ ms (c), $T_{60} = 800$ ms (d), and $T_{60} = 1000$ ms (e), respectively.

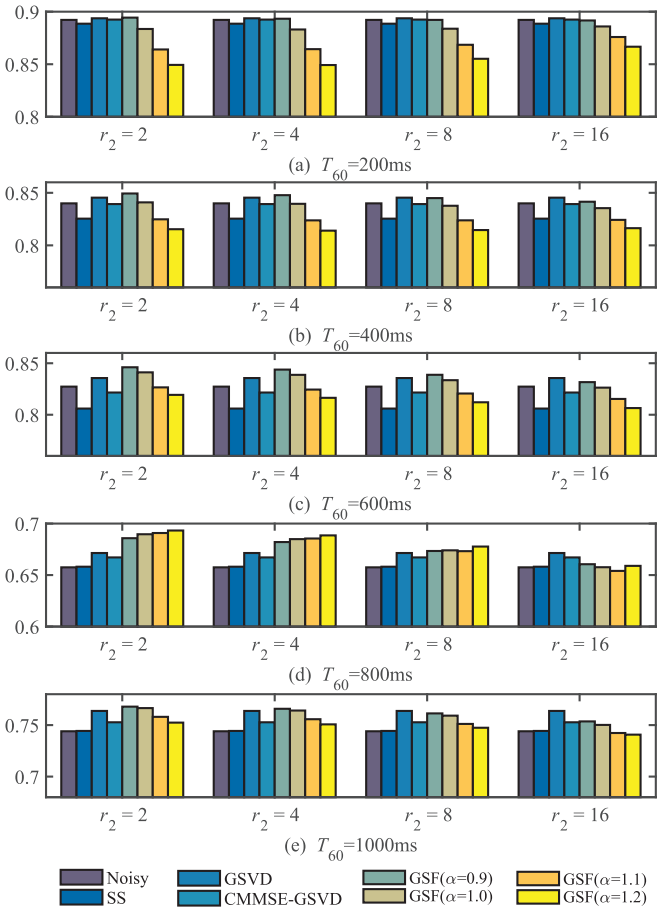

Fig. 5. STOI scores for Rooms A to E with T_{60} = 200 ms (a), T_{60} = 400 ms (b), T_{60} = 600 ms (c), T_{60} = 800 ms (d), and T_{60} = 1000 ms (e), respectively.

on the PESQ scores in Fig. 3, where only the CMMSE-GSVD method is comparable with the proposed GSF method when $T_{60} \geqslant 600$ ms. Among the three competing dereverberation methods, the CMMSE-GSVD method has the highest SRMR scores, while SS has the lowest SRMR scores in most cases. From Fig. 4, one gets that $r_2 = 2$ and $\alpha = 1.1$ can be chosen for moderate performance requirements except for $T_{60} = 200$ ms in practice.

4.3. STOI scores

The STOI scores are presented in Fig. 5. It is interesting to see that the STOI scores of the reverberant speech signals are even higher than those of corresponding dereverberated speech signals using SS, especially when $T_{60} \leqslant 600$ ms. This phenomenon could also be found in [42, Fig. 5], where the speech signals dereverberated by the method in [8] have much lower STOI scores than the reverberant speech signals. By properly choosing r_2 and α, the proposed method can achieve higher STOI scores in most cases when compared with the competing methods.

4.4. Discussions

Objective comparison results indicate that the proposed GSF method is promising in single-channel speech dereverberation. Compared with GSVD and CMMSE-GSVD, GSF has much less computational load and dose not require any late reverberation spectral variance estimation schemes. Moreover, GSF has the same computational complexity with SS, where both of them has an $O(1)$ time implementation for each time and frequency bin. However, SS still requires to estimate the late reverberation spectral variance for subtraction.

5. Conclusions

Motivated by guided imaging filtering method, we proposed a computationally efficient and effective method for speech dereverberation by using guided spectrogram filtering. It is interesting to see that the gain function can be computed from the normalized spectrogram of the reverberant speech signal and its enhanced version directly without estimating the late reverberant spectral variance. Performance evaluation shows that the proposed method for dereverberation has better performances in terms of both PESQ scores and SRMR scores. The proposed method can be easily extended to noise reduction when properly choosing the parameters in Algorithm 1, which could be one of future works. Another future work could concentrate on multi-channel speech dereverberation using guided spectrogram filtering.

Acknowledgements

This work was supported by National Science Fund of China Under Grant No. 61571435.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.apacoust.2017.11.016.

References

[1] Benesty J, Makino S, Chen J. Speech enhancement. Berlin: Springer-Verlag; 2005.
[2] Loizou PC. Speech enhancement: theory and practice. 2nd ed. Boca Roton, Florida: CRC PRESS; 2013.
[3] Benesty J, Sondhi MM, Huang Y. Springer handbook of speech processing. Berlin: Springer-Verlag; 2007.
[4] Naylor PA, Gaubitch ND. Speech dereverberation. London: Spinger-Verlag; 2010.
[5] Yegnanarayana B, Murthy PS. Enhancement of reverberant speech using LP residual signal. IEEE Trans Speech Audio Process 2000;8:267–81.
[6] Doclo S, Moonen M. GSVD-based optimal filtering for single and multimicrophone speech enhancement. IEEE Trans Signal Process 2002;50:2230–44.
[7] Lebart K, Boucher JM, Denbigh PN. A new method based on spectral subtraction for speech dereverberation. Acta Acust United Ac 2001;87:359–66.
[8] Wu M, Wang D. A two-stage algorithm for one-microphone reverberant speech enhancement. IEEE Trans Audio, Speech, Lang Process 2006;14:774–84.
[9] Habets EAP. Single- and multi-microphone speech dereveration using spectral enhancement. PhD Thesis, Technische Universiteit Eindhoven, The Netherlands, Jun. 25; 2007.
[10] Habets EAP, Gannot S, Cohen I. Late reverberant spectral variance estimation based on a statistical model. IEEE Signal Process Lett 2009;16:770–3.
[11] Kinoshita K, Delcroix M, Nakatani T, Miyoshi M. Suppression of late reverberation effect on speech using long-term multiple-step linear prediction. IEEE Trans Audio, Speech, Lang Process 2009;17:1–12.
[12] Erkelens JS, Heusdens R. Single-microphone late-reverberation suppression in noisy speech by exploiting long-term correlation in the DFT domain. In: Proc IEEE int conf audio, speech, and signal process, vol. 1; 2009. p. 3997–4000.
[13] Doclo S. Multimicrophone noise reduction and dereverberation techniques for speech applications. PhD dissertation, Dept Elect Eng, Katholieke Univ Leuven, Leuven, Belgium; May 2003.
[14] Mosayyebpour S, Esmaeili M, Gulliver TA. Single-microphone early and late reverberation suppression in noisy speech. IEEE Trans Audio, Speech, Lang Process 2013;21:322–35.
[15] Chen Z, Wang R, Yin F, Wang B, Peng W. Speech dereverberation method based on spectral subtraction and spectral line enhancement. Appl Acoust 2016;112:201–10.
[16] Fang Y, Feng H, Chen Y. A robust interaural time differences estimation and dereverberation algorithm based on the coherence function. Appl Acoust 2018;129:126–34.
[17] Zheng C, Peng R, Li X. A constrained MMSE LP residual estimator for speech dereverberation in noisy environments. IEEE Signal Process Lett 2014;21:1462–6.
[18] Tomasi C, Manduchi R. Bilateral Filtering for Gray and Color Images. In: Proc IEEE int conf. computer vision; 1998.
[19] Yang Q, Tan KH, Ahuja N. Real-time O(1) bilateral filtering. In: Proc IEEE conf computer vision and pattern recognition; 2009. p. 557–64.
[20] He K, Sun J, Tang X. Single image haze removal using dark channel prior. IEEE Trans Pattern Anal Mach Intell 2011;33:2341–53.
[21] He K, Sun J, Tang X. Guided image filtering. IEEE Trans Pattern Anal Mach Intell 2013;35:1397–409.
[22] Pinkowski B. Principal component analysis of speech spectrogram images. Pattern Recognit 1997;30:777–87.
[23] Nelson D, Cristobal G, Kober V, Cakrak F, Loughlin P, Cohen L. Denoising using

time-frequency and image processing methods. In: Proc SPIE 3807, advanced signal process algorithms, architectures, and implementation IX, 564, Nov. 2; 1999.
[24] Asaki K, Ogawa A. Reduction of noise in speech signals through image processing using the spectrogram. IEEJ Trans Electron, Inf Syst 2006;126:1483–9.
[25] Michelsanti D, Tan ZH. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. In: Interspeech 2017, Stockholm, Sweden.
[26] Xu H, Tan ZH, Dalsgaard P, Linderg B. Robust speech recognition by nonlocal means denoising processing. IEEE Signal Process Lett 2008;15:701–4.
[27] Mallawaarachchi A, Ong SH, Chitre M, Taylor E. Spectrogram denoising and automated extraction of the fundamental frequency variation of dolphi whistles. J Acoust Soc Am 2008;124:1159–70.
[28] Hu G, Wang DL. An auditory scene analysis approach to monaural speech segregation. In: Hansler E, Schmidt G, editors. Topics in acoustic echo and noise control. Heidelberg: Springer-Verlag; 2006. p. 485–515.
[29] Dennis J, Tran HD, Li H. Spectrogram image feature for sound event classification in mismatched conditions. IEEE Signal Process Lett 2011;18:130–3.
[30] Ajmera PK, Jadhav DV, Holambe RS. Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram. Pattern Recogn 2011;44:2749–59.
[31] Bradley JS, Sato H, Picard M. On the important of early reflections for speech in rooms. J Acoust Soc Am 2003;113:3233–44.
[32] Hu Y, Kokkinakis K. Effect of early and late refections on intelligibility of reverberated speech by cochlear implant listeners. J Acoust Soc Am 2014;135:EL22–8.
[33] Garofolo JS. Getting started with the DARPA TIMIT CD-ROM: an acoustic-phonetic continous speech database, Nat. Inst. of Standards and Technology (NIST), Gaithersburg, MD; 1993.
[34] Allen JB, Berkley DA. Image method for efficiently simulating small-room acoustics. J Acoust Soc Am 1979;65:943–50.
[35] Schwarz A, Kellermann W. Coherent-to-diffuse power ratio estimation for dereverberation. IEEE/ACM Trans. Audio, Speech, and Lang Process 2015;23:1006–18.
[36] http://reverb2014.dereverberation.com/download.html.
[37] Kinoshita K, Delcroix M, Yoshioka T, Habets EAP, H-Umbach R, Leutnant V, et al. Summary of the REVERB challenge. In: The REVERB workshop in Int Conf on Audio, Speech, and Lang Process, Florance, Italy, May 10; 2014. http://reverb2014.dereverberation.com/workshop/slides/reverb-summary.pdf.
[38] Falk TH, Zheng C, Chan WY. A non-intrusive quality and intellgibility measure of reverberant and dereverberated speech. IEEE Trans Audio, Speech, Lang Process 2010;18:1766–74.
[39] ITU-T P.862, "Perceptual evaluation of speech quality: an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. In: Int Telecom Union; 2001.
[40] Taal CH, Hendriks RC, Heusdens R, Jensen J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: Proc inter conf on acoust speech and signal process (ICASSP); 2010. p. 4214-7.
[41] Taal CH, Hendriks RC, Heusdens R, Jensen J. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans Audio, Speech, Lang Process 2011;19:2125–36.
[42] Han K, Wang Y, Wang D, Woods WS, Merks I, Zhang T. Learning spectral mapping for speech dereverberation and denoising. IEEE/ACM Trans Audio, Speech, Lang Process 2015;23:982–92.