Guided spectrogram filtering for speech dereverberation

Chengshi Zheng, Zheng-Hua Tan, Renhua Peng, Xiaodong Li

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

\[ x(n) = \sum_{m=-\infty}^{\infty} h(n,m)s(m), \]

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
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

\[
x(n) = \sum_{m=-D_0}^{n} h(n,m)s(m) + \sum_{m=-D_0}^{n-D_0} h(n,m)s(m),
\]

where \( D_0 \) denotes the filter length of early reflections. (2) is further written as

\[
x(n) = z_x(n) + z_x(n),
\]

where \( z_x(n) = \sum_{m=-n}^{n-D_0} h(n,m)s(m) \) denotes the early speech component and \( z_x(n) = \sum_{m=-n-D_0}^{n-D_0} h(n,m)s(m) \) 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

\[
X(k,l) = Z_x(k,l) + Z_x(k,l),
\]

where \( X(k,l) = \sum_{n=0}^{N-1} x(n + lR)w(n)e^{-j2\pi nk} \) is the STFT of \( x(n) \), and \( k \) and \( l \) are, respectively, the frequency bin index and the frame index. \( R \) is the frame length and \( R \) is the frame shift. Analogous to \( X(k,l),Z_x(k,l) \) and \( Z_x(k,l) \) correspond to the STFTs of \( z_x(n) \) and \( z_x(n) \), respectively. Spectrograms of the microphone signal, the early speech component, and the late speech component are denoted as \( \mathcal{X}(k,l) = 20\log_{10}(|X(k,l)|)^2 \), \( z_x(k,l) = 20\log_{10}(|Z_x(k,l)|)^2 \) and \( z_x(k,l) = 20\log_{10}(|Z_x(k,l)|)^2 \), 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 \( \mathcal{X}(k,l) \) as an example, we have

\[
\tilde{\mathcal{X}}(k,l) = \frac{\mathcal{X}(k,l)}{\max[\mathcal{X}]},
\]

where \( \max[\mathcal{X}] \) extracts the maximum value of the matrix \( \mathcal{X} \). The element of \( \tilde{\mathcal{X}} \) is given by

\[
\tilde{\mathcal{X}}(k,l) = \begin{cases} \mathcal{X}(k,l), & \text{if } \mathcal{X}(k,l) > \min[\tilde{\mathcal{X}}] \\ \min[\tilde{\mathcal{X}}], & \text{otherwise} \end{cases}
\]

where \( \min[\tilde{\mathcal{X}}] = \max[\mathcal{X}] - 255 \).

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 be 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 \( z_x(n) \) from the microphone signal \( x(n) \).

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 \( \tilde{\mathcal{X}} \). The element of the filtering input spectrogram \( \tilde{\mathcal{P}} \) can be computed from \( \tilde{\mathcal{X}} \) directly, which is given by

\[
\tilde{\mathcal{P}}(k,l) = \beta \tilde{\mathcal{X}}(k,l)+1-(1-\beta)\tilde{\mathcal{X}}(k,l),
\]

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

\[
\tilde{Q} = A\tilde{\mathcal{X}} + B,
\]

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

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

\[
\sigma_{\tilde{\mathcal{X}}}^2(k,l) = m_{\tilde{\mathcal{X}}}(k,l)(m_{\tilde{\mathcal{X}}}(k,l))^2,
\]

where

\[
m_{\tilde{\mathcal{X}}}(k,l) = \sum_{l-l_2}^{l+l_2} \sum_{k-k_2}^{k+k_2} \tilde{\mathcal{X}}(k,l)/N,
\]

and

\[
m_{\tilde{\mathcal{X}}}(k,l) = \sum_{l-l_2}^{l+l_2} \sum_{k-k_2}^{k+k_2} \tilde{\mathcal{X}}(k,l)/N,
\]

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 \( \tilde{\mathcal{X}} \). 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 \( \tilde{\mathcal{X}} \) and the filtering input spectrogram \( \tilde{\mathcal{P}} \) at each frequency bin \( k \) and \( l \), given by

\[
cov_{\tilde{\mathcal{X}}\tilde{\mathcal{P}}}(k,l) = m_{\tilde{\mathcal{X}}\tilde{\mathcal{P}}}(k,l) - m_{\tilde{\mathcal{X}}}(k,l)m_{\tilde{\mathcal{P}}}(k,l),
\]

where...
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 \( \tilde{X}(k,l) \)
   a) Apply the STFT to obtain \( X(k,l) \)
   b) Use (5)–(7) to calculate \( \tilde{X}(k,l) \)

2. **Guided Filter**
   **Input:** the filtering input spectrogram \( \tilde{X}(k,l) \) and \( \epsilon \)
   **Output:** the filtering output spectrogram \( \tilde{Q}(k,l) \)
   a) Compute \( \tilde{P} \) using (8)
   b) Compute \( A \) and \( B \) using (10)–(19)
   c) Compute \( \tilde{Q} \) using (9)

3. **Speech Reconstruction**
   **Input:** \( \tilde{X}(k,l), \tilde{Q}(k,l) \) and \( X(k,l) \)
   **Output:** the time-domain enhanced speech \( \tilde{z}_e(n) \)
   a) Compute the enhanced spectrogram \( \tilde{G}(k,l) \) using (20)
   b) Normalize the enhanced spectrogram using (21)
   c) Calculate the gain function using (22) for dereverberation
   d) Reconstruct \( \tilde{z}_e(n) \) 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 \( r_1 = 1, r_2 = 2, \alpha = 1.0, G_{\text{min}} = -17 \) dB, \( K = 512, R = 128 \) and \( \epsilon = 0.8^2 \) 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,
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, \( \approx T_{60} \) 0.6 s
4. Room D: variable reverberation room, 4.5 m × 3.5 m × 3 m, \( \approx T_{60} \) 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, \( \eta = 1, K = 512, R = 128, \epsilon = 0.8^2 \), and \( G_{\mathrm{noise}} = -17 \) dB are fixed for the sampling rate \( f_s = 16,000 \) Hz, while both \( \eta \) and \( \epsilon \) are variables, where both of them are set to different values, i.e. \( \eta \in \{2, 4, 8, 16\} \) and \( \epsilon \in \{0.9, 1.0, 1.1, 1.2\} \).

4.1. PESQ scores

The PESQ scores are shown in Fig. 3. It is obvious that the performance of GSF is highly correlated with \( \eta, \alpha \) 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 \( \eta \), e.g., \( \eta = 2 \), while it is better to use a larger value of \( \eta \) when the reverberation time \( T_{60} \) is larger than 600 ms, e.g., \( \eta = 4 \) or \( \eta = 8 \). 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 \( \alpha \) 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, \( \alpha \) cannot be too large. PESQ scores significantly reduce when \( \alpha \) 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 \( T_{60} \leq 600 \) ms when \( \eta = 2 \) 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
on the PESQ scores in Fig. 3, where only the CMMSE-GSVD method is comparable with the proposed GSF method when \( T_{60} \geq 600 \) 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 \( T_{60} \leq 600 \) 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 \( \alpha \), 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 does 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

time-frequency and image processing methods. In: Proc SPIE 3807, advanced signal process algorithms, architectures, and implementation IX, 564, Nov. 2; 1999.


