Single-Channel Speech Dereverberation in Noisy Environment for Non-Orthogonal Signals*

Abdullah Fahim, Prasanga N. Samarasinghe, Thushara D. Abhayapala
Research School of Engineering, Australian National University.

Summary

The detrimental effect of speech reverberation reduces speech quality, limits the performance of automatic speech recognition systems and impairs hearing aids. Spectral enhancement (SE) is a popular method for suppressing the late reverberation and background noise. However, conventional SE-based approaches assume orthogonality between the desired and undesired signal components. This orthogonality assumption does not hold true in most of the practical cases due to a limited time-domain support and the short-time stationarity of the speech signals, and thereby, affects estimation accuracy. To circumvent this issue, Lu et al. relaxed the orthogonality assumption by proposing a geometric approach to spectral subtraction (GSS) and evaluated their algorithm against different kinds of background noise. In our work, we comprehensively analyze the model by virtue of a simplified GSS transfer function to gain an insight into the algorithm. We conduct a series of experiments to validate GSS and explore its limitations in diverse realistic scenarios with both reverberation and background noise through a comprehensive end-to-end system for speech dereverberation and noise suppression. We also analyze the performance of GSS using the experimental data of the 2014 REVERB challenge and compare it with other conventional approaches such as spectral subtraction, Wiener Filter, minimum mean square error short-time spectral amplitude estimator and log spectral amplitude estimator, as well as with the contemporary methods of the 2014 REVERB challenge.

1 Introduction

Speech reverberation is the distortion of the sound by its delayed and attenuated copies, originated from the reflections of surrounding walls or objects. This phenomenon of reverberation reduces speech intelligibility and degrades the performance of hearing aids and automatic speech recognition (ASR) systems [1, 2]. The resulted speech distortion can be contributed by two distinct components of reverberation - the early reflections which cause coloration of speech, and the late reflections that contribute to echo and other significant distortions. While both types of reverberation components cause speech deformation, it is the latter one that is found to be more detrimental in practice [3, 4]. In real environments, reverberation is often accompanied with background noise. Thus it is important to perform speech dereverberation and noise suppression at the same time. In our current work, we design a combined speech dereverberation and noise suppression system using a modified spectral subtraction (SS) method to address the non-orthogonality between the signal components and evaluated and compared the performance with other conventional and contemporary methods.

1.1 Literature Review

In the past, researchers had offered diverse solutions to reduce speech reverberation. However, each dereverberation method exhibits its limitation in specific scenarios. Inverse filtering is a common dereverberation technique where a suitable filter is designed to nullify the room effect after estimating the room transfer function (RTF) [5]. A two-stage dereverberation algorithm was proposed in [6], which involved room impulse response (RIR) estimation from the null subspace of the data matrix and equalization of microphone signals using multi-channel inverse theorem (MINT) [7]. However, the estimation of the RIRs is a challenging task and the MINT algorithm's performance was found extremely sensitive to the RIR estimation error. In [8], Schwartz et al. proposed an online recursive expectation-maximization scheme to estimate the clean speech signal and the acoustic system simultaneously. Yoshioka et al. modeled the acoustic path as an auto-regressive system and used an all-pole model of the speech signal to perform noise suppression and dereverberation with a multi-channel Wiener filter (WF) [9] assuming a known noise power spectral density (PSD).

Supervised learning-based dereverberation meth-

 Published in Acta Acustica united with Acustica, Volume 104, Number 6. DOI: https://doi.org/10.3813/AAA.919270

ods extract and exploit speech feature vectors to suppress reverberation [10–13]. The process generally involves the mapping of speech feature vector during training using machine learning algorithms. During test, the estimated mapping from training sequences is used to achieve speech enhancement. However, the learning-based algorithms are generally resource intensive and require a long time context, which makes them hard to implement in real time processing. In [14], Nakatani et al. proposed harmonicity-based dereverberation (HERB) methods, which modeled RIR inverse filters as a ratio of the direct path component to the received signal. The design of the inverse filters exploited the harmonicity characteristics of the speech signal and estimated the filter coefficients in two distinct methods - one method estimated the average filter that transformed reverberant signals into harmonic signals, while the other method used a minimum mean squared error criterion that evaluated the quasi-periodicity of target signals. HERB algorithms take relatively longer time to converge, which also makes them difficult to use in real time processing. Linear predictive multi-input equalization (LIME) algorithm was used in [15] to achieve muti-channel dereverberation. The whitened speech residuals from the LIME output was mixed with the estimation of source auto regressive polynomials to obtain clean speech. In [16], the authors proposed an alteration arguing that any common zero in the RTFs of the multi-channel system degrades the LIME algorithm's performance.

Spectral subtraction is another class of dereverberation techniques, popular for its simplicity and low computational cost. In 2001, Lebart et al. formulated the use of SS for speech dereverberation [17] using Polack's statistical model of RIR [18]. Habets enhanced the method for single-channel and multi-channel cases through a series of publications [19–22]. A long-term multi-step linear prediction-based late reverberation signal estimation was used in SS by Kinoshita et al. in [2]. Wisdom et al. proposed speech coherence-based minimum mean square error (MMSE) log spectral amplitude estimator in [23]. Another variation of SS-based method was proposed by Cauchi et al. who incorporated temporal cepstrum smoothing [24]. Wu et al. estimated the late reverberation power spectrum using an asymmetrical smoothing window based on Rayleigh distribution [25]. Veras et al. extended Wu's method in their formulation of speech dereverberation in [26]. Kokkinakis et al. used variable subtraction factor as a function of the *a posteriori* signal to noise ratio (SNR) and evaluated the performance in cochlear implant devices [27]. However, most of the spectral enhancement techniques assume that the speech signal is orthogonal to the undesired signal, be it a random background noise or reverberation, and ignore any cross-term between the signal components. However, Lu et al. argued that the cross-term was not necessarily zero in all the scenarios and depended on the *a priori* SNR in practical cases involving white background noise [28]. They took a geometric approach to the spectral subtraction (GSS) that does not assume the signal orthogonality and demonstrated the impact for background noise suppression.

1.2 Our Contribution

The primary objective of this work is to gain a theoretical insight into GSS algorithm and examine its performance in a realistic reverberant environment. We discuss different theoretical aspects of GSS by simplifying the GSS transfer function proposed by Lu et al. [28] and examine its basic limitations. The original evaluation of GSS [28] assumed a non-reverberant environment, we extend that to evaluate GSS performance in a more realistic scenario by conducting a series of experiments in different reverberant and noisy environments. We prove that GSS produces better results with an accurate PSD knowledge, however, its performance is very sensitive to PSD estimation error. The impact of the decision-directed approach in *a priori* signal to noise ratio (SNR) estimation is assessed where we find that, contrary to the usual cases [28,32], GSS performs better with a lower smoothing constant. The performance of GSS is analyzed and compared with different conventional approaches using oracle PSD knowledge to isolate the true improvement offered by GSS. This work also presents performance evaluation of GSS with blind PSD estimation and compares the results with 4 conventional methods as well as with 12 contemporary methods from the 2014 REVERB challenge speech enhancement (RCSE2014) task. The latter experiments with estimated PSDs use RCSE2014 dataset which evaluates performances in 6 different environments using 2176 speech signals. In doing so, we design an end-to-end system for blind speech dereverberation and noise suppression where no prior knowledge of the room or the source characteristics are required.

The paper outlines as follows. Section 2 contains the problem statement. The limitations of conventional approaches are discussed in Section 3. We also present an overview of GSS and discuss its limitation. In Section 4, we describe an end-to-end system model for blind speech dereverberation and noise suppression to evaluate the performance of GSS. Finally, we present our experimental results in Section 5 in different reverberant conditions and compare it with RCSE2014 result set.

2 Problem formulation

A speech signal $y(n)$ captured in a distant microphone is described as

$$y(n) = \underbrace{s(n) * h(n)}_{x(n)} + v(n) \tag{1}$$

where $*$ denotes convolution, $s(n)$ is the clean speech signal, $h(n)$ represents the RIR, $v(n)$ is the background noise, and $x(n)$ is defined as the noise-suppressed reverberant signal. The RIR can be interpreted as a composition of two components $h_e(n)$ and $h_r(n)$, which correspond to early and late reverberation, respectively. Hence, we define

$$h_e(n) = \begin{cases} h(n) & \text{for } 0 \leq n < N_e, \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$h_r(n) = \begin{cases} h(n) & \text{for } n \geq N_e, \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where N_e separates the early reflections from the late reverberation. By substituting (2) and (3) into (1), we derive

$$\begin{aligned} y(n) &= s(n) * h_e(n) + s(n) * h_r(n) + v(n) \\ &= \underbrace{\sum_{m=n-N_e+1}^{n} s(m)h(n-m) +}_{x_e(n)} \\ &\quad \underbrace{\sum_{m=-\infty}^{n-N_e} s(m)h(n-m) + v(n)}_{x_r(n)} \end{aligned} \tag{4}$$

where $x_e(n)$ represents the direct signal and few early reflections, whereas $x_r(n)$ is the late reverberation component of the signal. In a two-stage algorithm, we first suppress the late reverberation component and recover the noisy dereverberated speech. Subsequently, we use the noisy dereverberated speech as the input to a denoising block to remove the background noise. For this purpose, we define the noisy dereverberated signal as

$$x_d(n) = x_e(n) + v(n). \tag{5}$$

Therefore, (4) takes the form of

$$y(n) = x_d(n) + x_r(n). \tag{6}$$

Our objective is, to estimate the clean speech $x_e(n)$ given the received signal $y(n)$. In doing so, we first estimate the spectral magnitude $|X_d(\omega)|$ of $x_d(n)$ and subsequently recover the spectral magnitude of the clean speech $|X_e(\omega)|$ using $|X_d(\omega)|$, where $X_d(\omega)$ and $X_e(\omega)$ are the Fourier transform (FT) coefficients of $x_d(n)$ and $x_e(n)$, respectively. Finally, we use the noisy phase with the estimation of $|X_e(\omega)|$ to construct the signal $x_e(n)$.

3 Spectral enhancement

3.1 The Conventional Approaches

For the theoretical analysis, we shall consider the generalized noise suppression model of (1), however, the discussion is equally applicable for the dereverberation technique where the late reverberation component is modeled as an additive interference.

3.1.1 Conventional Spectral Subtraction

The conventional SS is based on the instantaneous signal spectra. From (1), the squared-magnitude spectrum of $y(n)$ is give by

$$\begin{aligned} |Y(\omega)|^2 = |X(\omega)|^2 + |V(\omega)|^2 + \\ \underbrace{2\,|X(\omega)|\,|V(\omega)|\,\cos\left(\theta_{XV}(\omega)\right)}_{\text{Cross-terms}} \end{aligned} \tag{7}$$

where $\{Y(\omega), X(\omega), V(\omega)\}$ are the FTs of $\{y(n), x(n), v(n)\}$, $\theta_{XV}(\omega)$ is the phase difference between $X(\omega)$ and $V(\omega)$, and $|\cdot|$ denotes absolute value. If $X(\omega)$ and $V(\omega)$ are orthogonal, i.e. $\theta_{XV} = \pi/2\ \forall\omega$, the cross-terms of (7) becomes zero and $|X(\omega)|$ can be estimated by

$$\begin{aligned} |\hat{X}(\omega)| &= \sqrt{|Y(\omega)|^2 - |V(\omega)|^2} \\ &= |Y(\omega)|\ \underbrace{\sqrt{1 - \frac{1}{\gamma(\omega)}}}_{H_{ss}(\omega)} \end{aligned} \tag{8}$$

where $\gamma(\omega) = \frac{|Y(\omega)|^2}{|V(\omega)|^2}$ is the *a posteriori* SNR based on the instantaneous signal spectra and $H_{ss}(\omega)$ is the SS gain function. The estimated signal magnitude is then combined with the phase of the noisy signal $Y(w)$ to construct the recovered signal $\hat{X}(\omega)$. It is worth noting that, the noisy phase is used in signal reconstruction due to the fact that the phase distortion is largely inaudible, validated by the evaluations of the perceptual effects of simulated phase distortions [29, 30].

3.1.2 Wiener Filter

In the Wiener filter theory, an optimum filter $H_w(\omega)$ is designed to estimate $x(n)$ by minimizing the mean square error of the estimation, and the solution takes the form of [29]

$$H_w(\omega) = \frac{S_{xy}(\omega)}{S_{yy}(\omega)} \tag{9}$$

where $S_{xy}(\omega)$ is the cross spectral density (CSD) of $x(n)$ and $y(n)$ whereas $S_{yy}(\omega)$ is the PSD of $y(n)$, with ω denoting the angular frequency. Taking the FT of the auto-correlation of (1) and the cross-correlation between $x(n)$ and (1), we get

$$S_{yy}(\omega) = S_{xx}(\omega) + S_{vv}(\omega) + S_{xv}(\omega) + S_{vx}(\omega) \tag{10}$$

$$S_{xy}(\omega) = S_{xx}(\omega) + S_{xv}(\omega) \tag{11}$$

where $S_{xx}(\omega)$ and $S_{vv}(\omega)$ are the PSDs of $x(n)$ and $v(n)$, respectively, and $S_{xv}(\omega)$ and $S_{vx}(\omega)$ are the CSDs between $x(n)$ and $v(n)$ (hereafter we shall denote these as the cross-terms). In the Wiener solution, the cross-terms are considered zero and hence, using (9), (10) and (11), the solution becomes

$$\begin{aligned}\hat{H}_w(\omega) &= \frac{S_{xx}(\omega)}{S_{xx}(\omega) + S_{vv}(\omega)} \\ &= \frac{\tilde{\xi}(\omega)}{\tilde{\xi}(\omega) + 1}\end{aligned} \tag{12}$$

where $\tilde{\xi}(\omega) = \frac{S_{xx}(\omega)}{S_{vv}(\omega)}$ is the *a priori* SNR.

3.2 Limitations of the Conventional Approaches

The basis for ignoring cross-terms in both conventional SS and Wiener solution is that the speech and noise signals are uncorrelated, i.e. $E\{X(\omega)V^*(\omega)\} = 0$ where $E\{\cdot\}$ denotes the expected value. However, though this assumption of uncorrelated speech and noise signals is reasonable and widely used in speech processing, it doesn't necessarily mean that the instantaneous signal spectra are orthogonal in each time-frequency bins. It is shown in [29] that as the averaging window increases, the time average of the estimated instantaneous power spectra converges to the time-averaged true power spectrum. However, due to the non-stationarity of the speech signal, it is not practical to take the average spectrum over a long window to avoid the smearing effect. Hence, the orthogonality assumption of (7) leads to overestimation or underestimation of the estimated signal depending on the actual value of $\theta_{XV}(\omega)$.

A similar issue affects the Wiener solution of (12) where the calculation of the PSD values requires the ensemble average of the signal spectra. In a practical implementation, the speech signal is assumed to be an ergodic process where the ensemble averaging is replaced with the time-averaging of the signal spectra. However, due to the aforementioned non-stationarity issue of the speech signal, only a small number of windows are used in time-averaging which causes significant deviation from the true expected values. The immediate effect of this deviation is that the assumption of zero cross-terms of (12) doesn't hold anymore which results in a significant signal estimation error.

In [28], Lu et al. showed that the relative cross-terms with respect to the noisy signal spectra are large around 0 dB *a priori* SNR and more severe as θ_{XV} approaches π. They further investigated the noisy signal with different types of background noise and observed that a significant amount of samples lies within the zone where the cross-terms are significant.

In order to investigate the impact of cross-terms on signal estimation, we define the cross-term estimation error differently from Lu et al. by relating it directly to the estimated signal as

$$\begin{aligned}\epsilon(\omega) &= \frac{\left||X(\omega)|^2 - |\hat{X}(\omega)|^2\right|}{|X(\omega)|^2} \\ &= \frac{2}{\sqrt{\xi(\omega)}}\ \left|\cos\left(\theta_{XV}(\omega)\right)\right|\end{aligned} \tag{13}$$

where $|X(\omega)|^2$ and $|\hat{X}(\omega)|^2$ are respectively defined in (7) and (8), and $\xi(\omega) = \frac{|X(\omega)|^2}{|V(\omega)|^2}$ is the *a priori* SNR based on the instantaneous signal spectra. From (13) it is clear that for non-orthogonal signal frames, the cross-term error maintains an inverse linear relation with the *a priori* SNR as shown in Fig. 1.

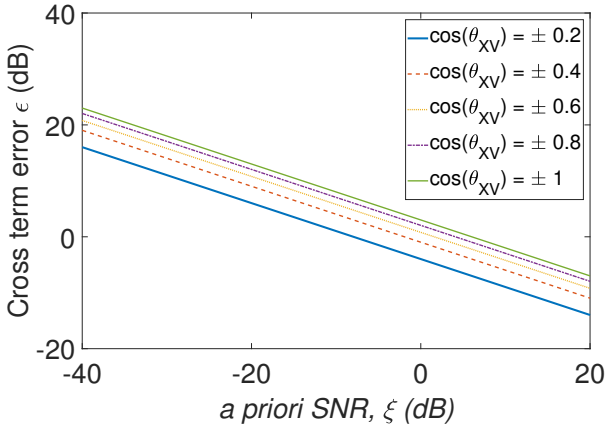


Figure 1: Cross-term error against *a priori* SNR for fixed phase differences between desired and undesired signal

To study the cross-term estimation error in practical environments, we measured the average mean square error in different noisy and reverberant environment with different short-time Fourier transform (STFT) window length. We define the average mean square cross-term estimation error as

$$\epsilon_{\text{avg}} = \frac{1}{L}\sum_{\forall \ell} \frac{\sum\limits_{\forall k}\left||X(\ell,k)|^2 - |\hat{X}(\ell,k)|^2\right|}{\sum\limits_{\forall k}|X(\ell,k)|^2} \tag{14}$$

$$= \frac{2}{L}\sum_{\forall \ell} \frac{\sum\limits_{\forall k}|X(\ell,k)|\ |V(\ell,k)|\ \left|\cos\left(\theta_{XV}(\ell,k)\right)\right|}{\sum\limits_{\forall k}|X(\ell,k)|^2} \tag{15}$$

where ℓ and k are the time and frequency bins of STFT, respectively, and L is the total number of time frames. Note that, Eq. (15) is obtained from (14) using (7) and (8). The results shown in Table 1 were obtained from the voiced-frames of 25 random speech signals assuming oracle PSD knowledge. The exclusion of the unvoiced frames was done to avoid error magnification in the frames where $X(\ell,k)$ is very small.

The results of Table 1 establish the presence of significant cross-term error irrespective of the STFT window length, especially at high reverberant/noisy condition. It is counter-intuitive that though the reverberant components share a higher degree of correlation with the desired speech signal compared to the white noise, it does not guarantee a higher cross-term error for reverberant component. This is due to the fact that the cross-term components originate from the lack of orthogonality of the desired and undesired speech components in the STFT time frame, even for the uncorrelated signals. The lack of orthogonality in this case is a direct result of limited time-domain support of the non-stationary speech signals, rather than the correlation between the signal components. It is important to note that, all the results of Table 1 are based on instantaneous spectra as used in conventional SS. The spectrogram of the cross-term estimation error for a random signal is shown in Fig. 2 for reference where the unvoiced parts of Fig. 2 were excluded while calculating the results in Table 1.

Table 1: Cross-term estimation error under different reverberation time (T_{60}), noise type, SNR, and window length of STFT.

	ϵ_{avg} (dB)			
Frame Size	8 ms	16 ms	32 ms	64 ms
T_{60}	Reverberant speech			
300 ms	-8.37	-8.5	-8.78	-8.96
600 ms	-2.9	-2.88	-2.81	-3.22
700 ms	-1.82	-1.77	-1.42	-2.44
SNR	Speech with air-condition noise			
10 dB	-4.66	-4.85	-4.69	-3.98
0 dB	0.34	0.15	0.31	1.02
SNR	Speech with white noise			
10 dB	-4.93	-4.81	-4.07	-3.65
0 dB	0.1	0.21	0.95	1.21

Hence, from the above discussion as well as the results presented in [28], we can conclude that the assumption of zero cross-terms results in a significant estimation error in the conventional spectral enhancement techniques.

3.3 Geometric Spectral Subtraction

To circumvent the cross-term estimation error, Lu et al. proposed an alternative approach for SS by taking a geometric approach [28]. In this section, we briefly discuss the GSS algorithm of [28].

The simplified signal model of (1) can be written

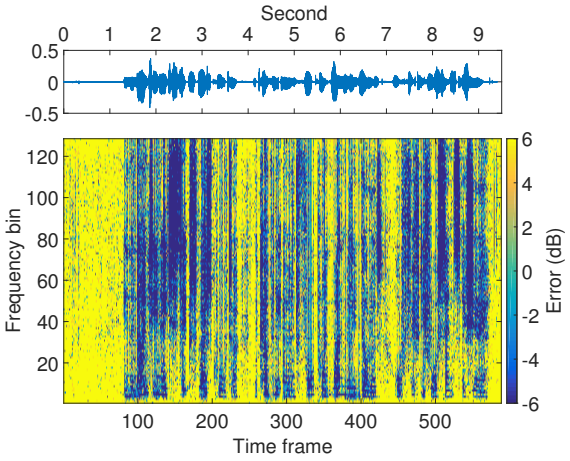


Figure 2: Cross-term error in a speech signal with 16 ms frames and no overlap, at 10 dB *a priori* SNR

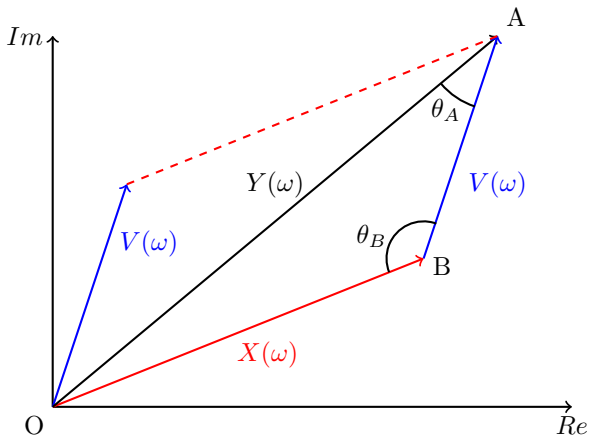


Figure 3: Phasor diagram of (16)

in FT domain as

$$Y(\omega) = X(\omega) + V(\omega). \tag{16}$$

The corresponding phasor diagram is drawn in Fig. 3. Using the law of Sines in the $\triangle OAB$ of Fig. 3, we get

$$\frac{|X(\omega)|}{|Y(\omega)|} = \frac{\sin(\theta_A)}{\sin(\theta_B)} = \sqrt{\frac{1-\cos^2(\theta_A)}{1-\cos^2(\theta_B)}}. \tag{17}$$

Also the laws of cosines in $\triangle OAB$ of Fig. 3 lead to

$$\cos(\theta_A) = \frac{|Y(\omega)|^2 + |V(\omega)|^2 - |X(\omega)|^2}{2|Y(\omega)||V(\omega)|} \tag{18}$$

$$\cos(\theta_B) = \frac{|X(\omega)|^2 + |V(\omega)|^2 - |Y(\omega)|^2}{2|X(\omega)||V(\omega)|}. \tag{19}$$

Using (17), (18) and (19), the gain function for GSS

is formulated as [28]

$$H_{GSS} = \frac{|X(\omega)|}{|Y(\omega)|} = \sqrt{\frac{1 - \frac{[\gamma(\omega)+1-\xi(\omega)]^2}{4\gamma(\omega)}}{1 - \frac{[\gamma(\omega)-1-\xi(\omega)]^2}{4\xi(\omega)}}} \quad (20)$$

where $H_{GSS} \in [0, 1]$ is enforced to avoid large estimation error [28].

3.4 Limitation of Geometric Spectral Subtraction

The GSS gain function of (20) requires both the *a priori* and *a posteriori* SNRs, similar to Ephraim and Malah's MMSE short-time spectral amplitude (STSA) estimator [32] and MMSE log spectral amplitude (LSA) estimator [33], however, differs in the sense that, it was formulated with the instantaneous signal spectra to calculate the SNRs, like the conventional SS. It is worth noting that, under the assumption of the zero cross-terms, (7) results in the following relation between $\xi(\omega)$ and $\gamma(\omega)$

$$\xi(\omega) = \gamma(\omega) - 1, \quad (21)$$

in which case the GSS gain function of (20) becomes identical to the conventional SS gain function of (8).

To have an insight of how GSS works, we can simplify the GSS gain function proposed in [28] as

$$H_{GSS} = \sqrt{\frac{\xi(\omega)}{\gamma(\omega)} \frac{4\gamma(\omega) - [\gamma(\omega) + 1 - \xi(\omega)]^2}{4\xi(\omega) - [\gamma(\omega) - 1 - \xi(\omega)]^2}} \quad (22)$$

$$= \sqrt{\frac{\xi(\omega)}{\gamma(\omega)}}, \text{ for } \xi(\omega) \neq \left[\sqrt{\gamma(\omega)} \pm 1\right]^2. \quad (23)$$

From (23), it is clear that, while GSS offers improvement by reducing cross-term error, the performance is limited by the estimation accuracy of $\xi(\omega)$. Furthermore, using the definition of ξ and γ in (23), one can observe that when the *a priori* and *a posteriori* SNRs are known precisely, GSS results in a perfect estimation of the desired signal. However, in reality, *a priori* SNR and the undesired PSD component are not readily available. While any error in undesired PSD estimation affects both GSS and conventional SS with an erroneous γ value, the estimation accuracy of *a priori* SNR determines the relative performance of GSS over conventional SS. A widely used technique for estimating *a priori* SNR is the decision-directed approach [32] where ξ is calculated as a weighted average of the past value and present *a posteriori* SNR. Therefore, the performance of GSS is largely affected by the choice of the weighting factor, as we will show in the results section. Hence, it is fair to conclude that while GSS offers an improvement on estimation accuracy by incorporating the cross-terms, its performance can suffer due to any estimation error of $\xi(\omega)$. Therefore, it is expected that, in scenarios that involve high cross-term components, GSS outperforms conventional SS.

A direct comparison between the GSS and WF is difficult to make as they are based on different underlying principles. GSS obviously gains the advantage of incorporating the cross-terms, but the comparative performance is also dictated by the different formulation of the *a priori* SNR. The GSS uses the instantaneous version $\xi(\omega)$ whereas the WF formulate $\tilde{\xi}(\omega)$ based on the statistical expectation. Hence, the improvements depend on how well $\xi(\omega)$ and $\tilde{\xi}(\omega)$ fit the *a priori* SNR estimation model, e.g. decision-directed approach [32]. As there is no easy way to make a theoretical comparison, we depend on the simulation results to compare the performances of GSS and WF.

4 System Model

In this work, we setup different experimental scenarios to evaluate and compare Lu's GSS [28] model against other conventional and contemporary methods. Unlike [28] where the evaluation took place for noise-only cases, we work with a more realistic scenario in a reverberant and noisy room. In the following sections, we describe different building blocks to design an end-to-end dereverberation and noise suppression system to be used in our experiments. As speech signal is considered short-time stationary, the whole process is performed in the STFT domain.

4.1 Dereverberation Block

Initially, we suppress the late reverberation component from the received signal (the motivation is discussed in Section 4.6). The late reverberation component is modeled as an additive interference in (6). Hence, comparing (6) with (1) and using (20), the dereverberated signal is

$$X_d(\ell, k) = Y(\ell, k) \underbrace{\sqrt{\frac{1 - \frac{[\gamma_d(\ell,k)+1-\xi_d(\ell,k)]^2}{4\gamma_d(\ell,k)}}{1 - \frac{[\gamma_d(\ell,k)-1-\xi_d(\ell,k)]^2}{4\xi_d(\ell,k)}}}}_{H_d(\ell,k)} \quad (24)$$

where *a priori* and *a posteriori* signal to reverberation ratios (SRR) are defined as

$$\xi_d(\ell, k) = \frac{|X_d(\ell, k)|^2}{|X_r(\ell, k)|^2}, \text{ and} \quad (25)$$

$$\gamma_d(\ell, k) = \frac{|Y(\ell, k)|^2}{|X_r(\ell, k)|^2}, \quad (26)$$

respectively, with $\{Y(\ell, k), X_d(\ell, k), X_r(\ell, k)\}$ being the STFTs of $\{y(n), x_d(n), x_r(n)\}$. Obviously, the

dereverberated signal energy $|X_d(\ell,k)|^2$ and the late reverberation energy $|X_r(\ell,k)|^2$ are unknown quantities. We estimate $|X_r(\ell,k)|^2$ from the received signal $Y(\ell,k)$, as described in the next section. The *a priori* SRR $\xi_d(\ell,k)$ is estimated using a decision-directed approach [32].

4.2 Late Reverberation Energy Estimator

4.2.1 Reverberation Time Estimator

The prerequisite for calculating the late reverberation energy is a good estimation of the reverberation time (RT), T_{60}. Several methods have been described in the literature for the blind estimation of RT [34–36]. In this work, we use the maximum likelihood based T_{60} estimator described in [36]. However, due to the frequency dependency of T_{60}, we first decompose the signal into 8 sub-bands using a 1/3 octave band filter bank. We then determine T_{60} for each sub-band using a maximum likelihood estimator and assign that value to the center frequency f_c of the sub-band. Subsequently, we use cubic spline interpolation to estimate T_{60} for each STFT frequency bin k.

4.2.2 Reverberation PSD Estimator

We use Lebart's late reverberation energy estimator [17] based on Polack's statistical model of RIR [18] to estimate reverberant PSD. Lebart et al. showed that the late reverberation energy is related to the observed reverberant signal from the past frames by [17]

$$S_{rr}(\ell,k) = e^{-2\Delta N_e}\; S_{xx}(\ell - \frac{N_e}{P}, k) \tag{27}$$

where $S_{rr}(\ell,k)$ is the the late reverberation PSD and P is the hop size of the STFT and

$$\Delta = \frac{3\;\log_e(10)}{T_{60}\;f_s} \tag{28}$$

with f_s being the sampling frequency. The statistical model of RIR can cause PSD estimation error in a realistic scenario due to the imperfect exponential envelope and deviation from the Gaussian distribution of a practical RIR [37]. However, from the implementation perspective, the statistical RIR model offers a simple and less resource intensive solution for late reverberation PSD estimation. A common practice for estimating the PSD component $S_{xx}(\ell,k)$ is by smoothing $|X(\ell,k)|^2$, however, in this case the noise-suppressed reverberant signal $X(\ell,k)$ is not available at this stage, hence we use the noisy reverberant signal to estimate $S_{xx}(\ell,k)$ as

$$\begin{aligned}\hat{S}_{xx}(\ell,k) &= S_{yy}(\ell,k)\\ &= \eta_x\; S_{yy}(\ell-1,k) + (1-\eta_x)\;|Y(\ell,k)|^2\end{aligned} \tag{29}$$

where $\eta_x \in [0,1]$ is a smoothing factor. A detail discussion on this estimation is presented in Section 4.6. Finally, we estimate the periodogram of the late reverberation component as

$$|\hat{X}_r(\ell,k)|^2 = S_{rr}(\ell,k). \tag{30}$$

4.3 Denoising Block

For noise suppression, we consider the signal model of (5) and estimate the desired signal magnitude using (20) as

$$\hat{X}_e(\ell,k) = \hat{X}_d(\ell,k)\underbrace{\sqrt{\frac{1-\frac{[\gamma_e(\ell,k)+1-\xi_e(\ell,k)]^2}{4\gamma_e(\ell,k)}}{1-\frac{[\gamma_e(\ell,k)-1-\xi_e(\ell,k)]^2}{4\xi_e(\ell,k)}}}}_{H_e(\ell,k)} \tag{31}$$

where $\hat{X}_d(\ell,k)$ is the estimated dereverberated signal from the dereverberation block and the *a priori* and *a posteriori* SNRs are defined as

$$\xi_e(\ell,k) = \frac{|\hat{X}_e(\ell,k)|^2}{|V(\ell,k)|^2}, \text{ and} \tag{32}$$

$$\gamma_e(\ell,k) = \frac{|\hat{X}_d(\ell,k)|^2}{|V(\ell,k)|^2}, \tag{33}$$

respectively, with $\{X_e(\ell,k), V(\ell,k)\}$ being the STFTs of $\{x_e(n), v(n)\}$. Similar to the dereverberation block, we have to estimate the noise energy $|V(\ell,k)|^2$ and use a decision-directed approach in estimating the *a priori* SNR $\xi_e(\ell,k)$.

4.4 Noise Energy Estimator

Assuming the background noise as a slowly time-varying process, we estimate the noise energy from the speech pauses. There are several VAD methods proposed in the literature. We use the method proposed by Verteletskaya et al. [38], where they determined the voice presence based on the signal periodicity, total voice band energy and the energy contribution of the higher frequency elements. Once we determine the voice presence in a frame, we calculate the noise energy using

$$|\hat{V}(\ell,k)|^2 = \eta_v\;|\hat{V}(\ell-1,k)|^2 + (1-\eta_v)\;|\hat{X}_d(\ell,k)|^2 \tag{34}$$

where $\eta_v \in [0,1]$ is a smoothing factor with the constraint of $\eta_v = 1$ in the voiced frames. It is worth mentioning that, in this work, we have not made a thorough investigation of the VAD performance, rather we design the VAD as an independent block which can be modified or replaced to the individual needs without impacting the overall design.

4.5 *A priori* and *a posteriori* SNRs

As both the GSS and conventional SS are based on the instantaneous signal spectra, the *a posteriori* SRR and SNR of (26) and (33) are formulated using the squared-magnitude spectra of the signal. However, in the simulation, we observed that an exponential smoothing results a better performance for both GSS and conventional SS-based algorithms. This can be a result of a reduced error variance due to the smoothing operation. Therefore, we use the following smoothed versions of the *a posteriori* SRR/SNR in the GSS gain functions of (24) and (31), respectively

$$\hat{\gamma}_d(\ell,k) = \beta\ \hat{\gamma}_d(\ell-1,k)+(1-\beta)\ \min\left\{\frac{|Y(\ell,k)|^2}{|\hat{X}_r(\ell,k)|^2}, \gamma_{mx}\right\} \tag{35}$$

$$\hat{\gamma}_e(\ell,k) = \beta\ \hat{\gamma}_e(\ell-1,k)+(1-\beta)\ \min\left\{\frac{|\hat{X}_d(\ell,k)|^2}{|\hat{V}(\ell,k)|^2}, \gamma_{mx}\right\} \tag{36}$$

where $\beta \in [0,1]$ is a smoothing constant, $\min\{\cdot\}$ operator takes the minimum of the variables inside and γ_{mx} represents the maximum allowable *a posteriori* SRR/SNR to avoid over-attenuation of the signal.

The estimation of the *a priori* SRR/SNR requires the knowledge of the processed signal energy. We adopt a widely used decision-directed approach, first proposed by Ephraim and Malah in [32], where the *a priori* SRR/SNR in a frame is estimated based on the previous frame's *a priori* SRR/SNR and current frames *a posteriori* SRR/SNR, respectively. In the original work, Ephraim et al. used (21) in the decision-directed approach by assuming zero cross-term. However, in GSS, the cross-terms are considered non-zero and (21) does not hold true. Therefore, we adopt Lu's modification to replace (21) by the theoretical minimum value of the *a priori* SRR/SNR [28] in the original decision-directed formulation and use the below *a priori* SRR/SNR in (24) and (31)

$$\hat{\xi}_d(\ell,k) = \max\Bigg\{\alpha\ \frac{|\hat{X}_d(\ell-1,k)|^2}{|\hat{X}_r(\ell-1,k)|^2}+ (1-\alpha)(\sqrt{\hat{\gamma}_d(\ell,k)}-1)^2,\ \xi_{min}\Bigg\} \tag{37}$$

$$\hat{\xi}_e(\ell,k) = \max\Bigg\{\alpha\ \frac{|\hat{X}_e(\ell-1,k)|^2}{|\hat{V}(\ell-1,k)|^2}+ (1-\alpha)(\sqrt{\hat{\gamma}_e(\ell,k)}-1)^2,\ \xi_{min}\Bigg\} \tag{38}$$

where $\alpha \in [0,1]$ is a smoothing constant, $\max\{\cdot\}$ operator takes the maximum of the variables inside and ξ_{min} represents the minimum allowable *a priori* SRR/SNR.

One important consideration of the SS-based solutions is that, they use the noisy phase with the estimated magnitude to reconstruct the signal. GSS also follows the same, with $H_d(\ell,k)$ in (24) and $H_e(\ell,k)$ in (31) are real valued quantities, $\hat{X}_d(\ell,k)$ and $\hat{X}_e(\ell,k)$ carry the same phase of noisy signal $Y(\ell,k)$. The use of the noisy phase for signal reconstruction is justified from the fact that the inaccurate phase information does not impact the overall SNR significantly [30], also the impact of the noisy phase is largely inaudible [29].

4.6 On Two-stage Approach of the Solution

In this work, we approach the solution using a two-stage algorithm. We perform the dereverberation task followed by the noise suppression. A two-stage approach is used for a better voice detection and noise PSD estimation from the received speech signal. As the late reverberation PSD overlaps with the signal PSD, the identification of the unvoiced frame can be erroneous and the PSD estimation accuracy can suffer. Hence, suppressing the late reverberation component before performing the noise PSD estimation, we expect to increase the estimation accuracy.

However, there exists a contradictory argument regarding the sequence of the two-stage algorithm, which suggests to perform the noise suppression ahead of the dereverberation task. The late reverberation energy estimator of (27) requires the noise-suppressed reverberant signal PSD $S_{xx}(\ell,k)$. If we perform the noise suppression at the end, $S_{xx}(\ell,k)$ remains unknown at the time of the dereverberation operation and the late reverberation PSD needs to be estimated from the noisy reverberant PSD $S_{yy}(\ell,k)$ instead of $S_{xx}(\ell,k)$. From (10) and (27), the late reverberation energy is

$$S_{rr}(\ell,k) = e^{-2\Delta N_e}\ \Bigg[S_{yy}(\ell-\frac{N_e}{L},k)- \underbrace{S_{xv}(\ell-\frac{N_e}{L},k) - S_{vx}(\ell-\frac{N_e}{L},k) - S_{vv}(\ell-\frac{N_e}{L},k)}_{\text{ignored terms}}\Bigg]. \tag{39}$$

Therefore, using $S_{yy}(\ell,k)$ to estimate $S_{rr}(\ell,k)$ would induce extra error due to the ignored terms shown in (39).

Hence, there has to be a compromise in determining the sequence of the dereverberation and noise suppression in a two-stage operation. This can be avoided in case an efficient noise PSD estimation is available under reverberant condition (or an efficient reverberant PSD estimation in a noisy environment). For this work, we chose the two-stage algorithm with "dereverberation-first" approach. Fig. 4 shows the

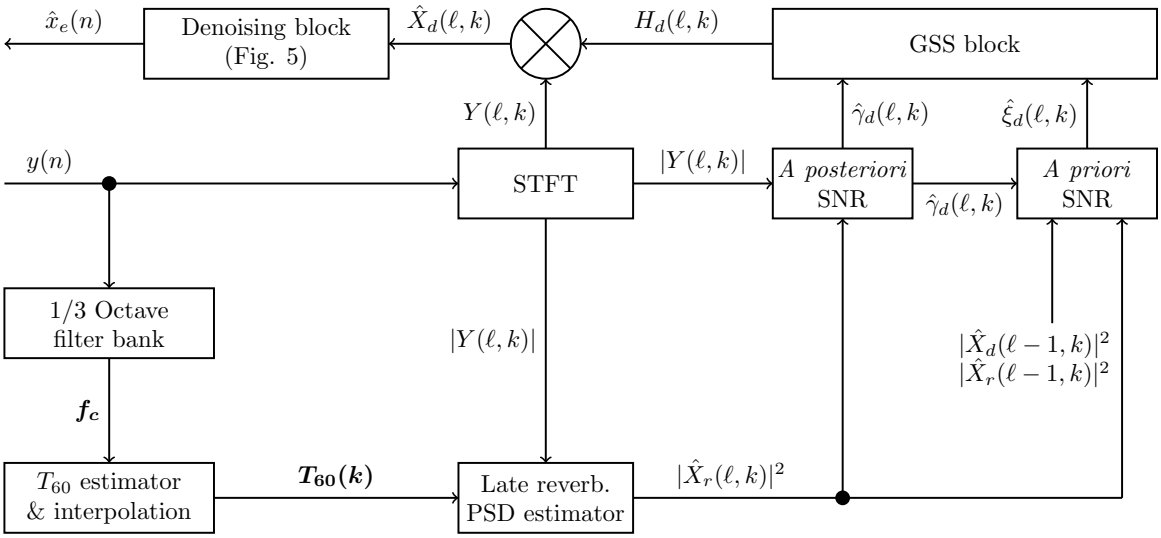


Figure 4: Block diagram of the system model.

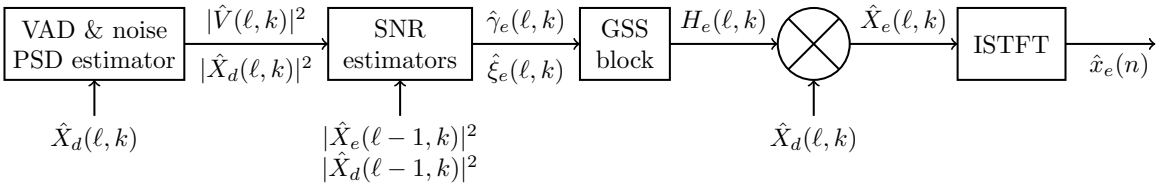


Figure 5: Denoising block. ISTFT stands for inverse STFT.

system model as a block diagram whereas the denoising block is shown separately in Fig. 5.

It should be noted that, irrespective of the sequence of the cascaded system, the second stage performance always suffers due to the estimation error at the first stage. In our design, any estimation error at the dereverberation block would affect the performance of the denoising block. As the gain functions of both dereverberation and denoising blocks are real, these estimation errors would affect the magnitude of the estimated quantity at each stage. In GSS, the estimation error in dereverberation stage results from two factors - inaccurate estimation of the late reverberation PSD and *a priori* SRR. Such an estimation error is unavoidable under the same underlying technique of GSS, and a compensation for such an error is very tough within the scope of GSS.

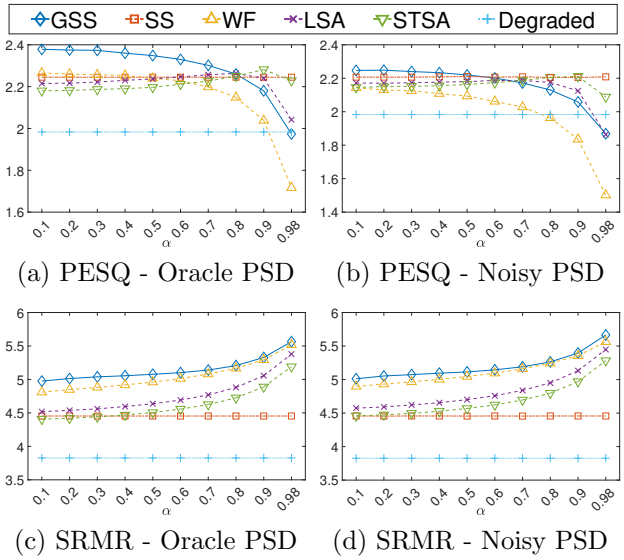
5 Experimental Results

5.1 RCSE2014 Evaluation Tool

The performance evaluation took place using the RCSE2014 [39] dataset. We used the official RCSE2014 evaluation tool (RCET) [39] for data generation and performance evaluation. The RCET contained 6 different RIRs measured under different room conditions which are listed in Table 2 for reference. Note that, the data of Table 2 were not used in the experiments as per RCSE2014 guideline. The dataset contains 362 speech files with $f_s = 16$ kHz at each reverberant room condition mixed with a background noise at 20 dB .

5.2 Parameters Settings & Evaluation Measures

We used a 16 ms window with 75% overlap for calculating a 256-point discrete FT. The parameter settings used in the experiments are chosen heuristically based on a subset of training data and shown in Table 3. We evaluated the performance by means of perceptual evaluation of speech quality (PESQ) [40], speech to reverberation modulation energy ratio (SRMR) [41], frequency-weighted segmental SNR (FWSegSNR), cepstral distance (CD) and log-likelihood ratio (LLR) [43]. For reference, higher PESQ, SRMR and FWSegSNR indicate a better performance, whereas the opposite is true for CD and LLR.



(a) PESQ - Oracle PSD (b) PESQ - Noisy PSD

(c) SRMR - Oracle PSD (d) SRMR - Noisy PSD

Figure 6: Performance of GSS with other conventional methods in terms of PESQ and SRMR for different α values with oracle PSD and noisy PSD estimation.

5.3 Selection of α

As discussed in Section 3.4, the accuracy of *a priori* SNR estimation plays a vital role on the performance of GSS. Most of the spectral enhancement techniques take a decision-directed approach [32] in calculating *a priori* SNR where the weighting factor α determines the bias of estimation towards past or present spectral cues (Eq. (37) and (38)). Historically, a near-unity value of α is believed to enhance the performance of the spectral subtraction based algorithms [32], which was also followed by Lu et al. [28]. However, while investigating the impact of α in GSS, we found that both dereverberation and denoising worked best with a mid-ranged α and suffered a drastic degradation in speech quality after $\alpha = 0.9$.

Figure 6 plots the PESQ and SRMR values for different methods against different α values under the assumption of a perfect PSD estimation (oracle PSD) and noisy PSD estimation (added white noise to oracle PSD at 10 dB SNR). It is interesting to note that, while PESQ shows a significant degradation at higher α, the SRMR improves consistently with α values. This indicates that at higher α, the dereverberation improves at the cost of signal quality. It can also be noted that with increasing PSD estimation error, the performance of GSS starts to degrade and beyond some threshold it becomes inferior compared to the conventional SS irrespective of the value of α. Note that, conventional SS does not depend on *a priori* SNR and hence, exhibits a constant performance over different α.

Among the other methods, STSA shows improved performance with increasing alpha, however suffers degradation at $\alpha = 0.98$ in noisy PSD estimation. The WF, which relies only on ξ (Eq. (12)) follows the similar trend as GSS, however, GSS always performs better due to the inclusion of the corrective *a posteriori* term γ. Note that, Fig. 6 shows the average results for 25 random speech signals from the training dataset and used to determine α for the evaluation dataset based on individual PESQ values. For the simulations in Section 5.4 and 5.5, we chose $\alpha = 0.9$ for STSA, $\alpha = 0.8$ for LSA, and $\alpha = 0.4$ for the remaining methods to ensure that each individual method exhibits the best PESQ in Fig. 6 for the corresponding α. Note that, we performed similar studies for η_x and β as we did for α, however, we did not observe any significant shift in the relative performances between GSS and other competing methods within the range of $\eta_x \in [0.8, 0.95]$ and $\beta \in [0.5, 0.9]$.

Table 2: Room geometry and RT

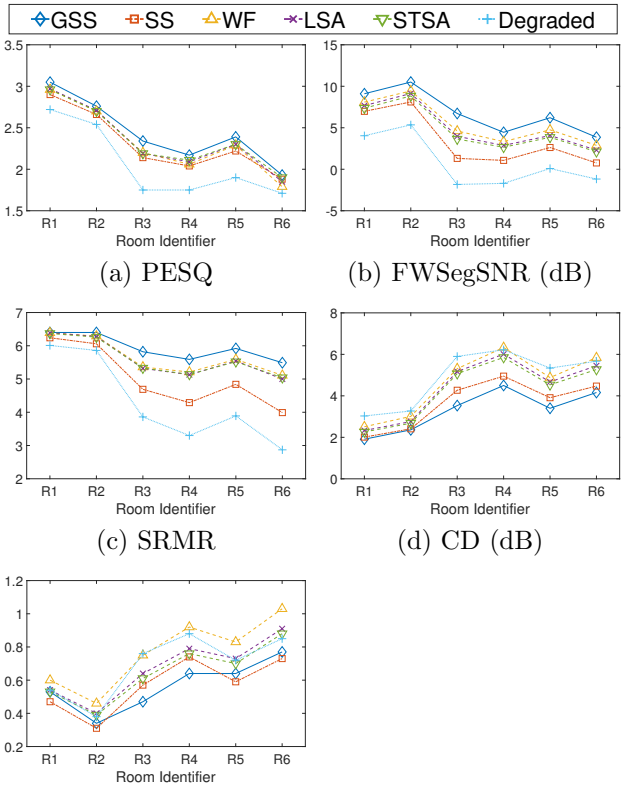
Identifier	Source to microphone distance	RT
R1	50 cm	0.3s
R2	200 cm	0.3s
R3	50 cm	0.6s
R4	200 cm	0.6s
R5	50 cm	0.7s
R6	200 cm	0.7s

Table 3: Parameter settings used in the simulation

Parameter	**Value**
α	See Section 5.3
β, η_v	0.6
η_x	0.85
γ_{mx}	13 dB
ξ_{min}	-26 dB
N_e	.05 f_s

5.4 Comparison with the Conventional Methods Using Oracle PSD

In our first comparative evaluation, we measured the performance of GSS with 4 different conventional methods. In this part of experiments, we considered oracle PSD knowledge to determine the true improvement of the comparing techniques unaffected by the PSD estimation accuracy. For the experiments, we used 25 random speeches in each reverberant condition of RCET with the recorded air-condition noise at 10 dB SNR. The oracle PSDs were computed using



(a) PESQ (b) FWSegSNR (dB)

(c) SRMR (d) CD (dB)

(d) LLR

Figure 7: Comparison of GSS with other conventional approaches for oracle PSD knowledge.

below exponential averaging formula

$$S_{rr}(\ell,k) = \eta\ S_{rr}(\ell-1,k) + (1-\eta)\ |X_r(\ell,k)|^2 \quad (40)$$

$$S_{vv}(\ell,k) = \eta\ S_{vv}(\ell-1,k) + (1-\eta)\ |V(\ell,k)|^2 \quad (41)$$

where $\eta = 0.85$ was used as the smoothing factor. Note that, GSS and SS were evaluated using instantaneous signal spectra, while the rests used the oracle PSD formula of (40) and (41).

The results shown in Fig. 7 clearly indicate that GSS consistently outperforms all other conventional methods at high level of PSD estimation accuracy. However, based on Fig. 6, the speech quality of GSS output is sensitive against PSD estimation error and can results in a significant speech distortion at a low level of PSD estimation accuracy. Hence, the expected accuracy of the PSD estimation block should determine the decision whether to prefer GSS over the conventional SS, as beyond a certain noise level, GSS fails to offer any improvement compared to the conventional SS.

In order to determine the computational complexity, we plot the average computation time for each methods in Fig. 8 which shows that the processing time of GSS is in the same order of the conventional SS and WF. The STSA and LSA took the longest time to process, mainly due to the requirements of the extra Bessel and exponential function blocks.

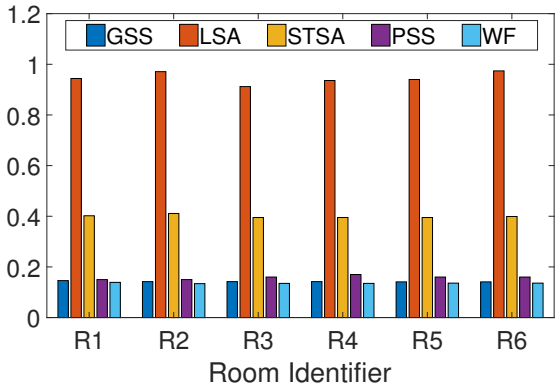
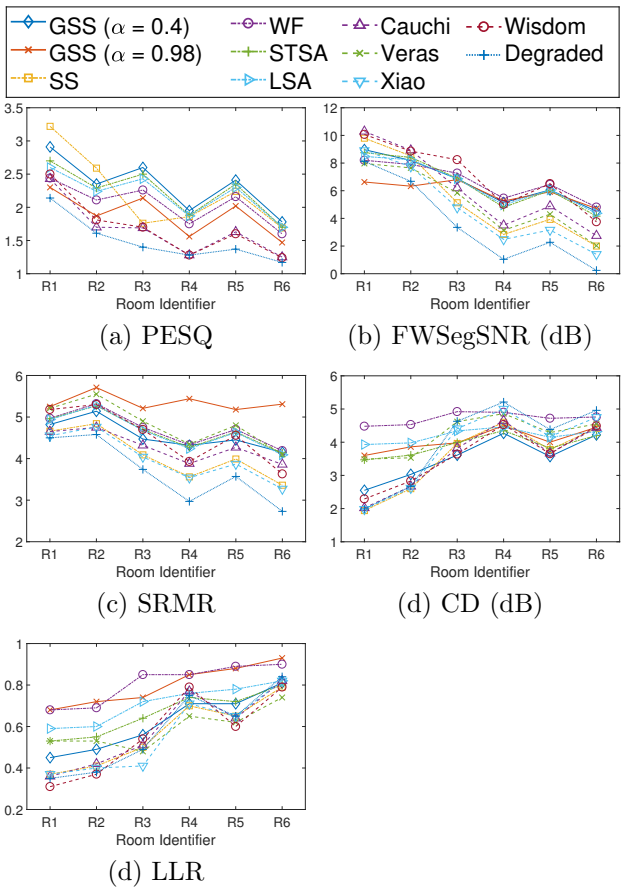


Figure 8: Average processing time per signal (assuming oracle PSD knowledge with 6.7s average signal duration).

5.5 Comparison with the conventional methods and RCSE2014 results based on estimated PSD

In this section, we compare the output of the proposed system with other contemporary single-channel methods from the RCSE2014 as well as with convention methods. We rigorously followed the instruction of RCSE2014 SE task and used the complete dataset to offer a true comparison with the RCSE2014 methods. Note that, while we used the same PSD estimator for GSS and other 4 conventional methods, we could not ensure the same PSD estimator for the RCSE2014 methods as we directly compared with the official results published in the workshop. However, the evaluation process of GSS used very basic and fundamental PSD estimators such as voice inactivity-based signal averaging for noise PSD and statistical RIR-based method for reverberation PSD. Both these methods have been in use for a long time, and it is fair to assume that the PSD estimators used in the RCSE2014 methods exhibited an improved, or at least a similar level of accuracy compared to those of the old techniques. We also include GSS performances based on two different values of α to show the impact of α on speech quality and dereverberation.

Fig. 9 shows the comparative performance of GSS and conventional methods with 4 SS-based methods [12, 23, 24, 26] from the RCSE2014 using a blind PSD estimation. Each RCSE2014 method in Fig. 9 is denoted by the name of the corresponding first author. The performance measures of the degraded speech are included in the figure for reference. Note that, as we update the noise PSD only during the speech pauses, we used a lower smoothing constant $\eta_v = 0.6$ to put extra confidence on the current frame estimate. However, we also used a spectral floor of 0.2 for the GSS gain function in the unvoiced frames to compensate any rapid fluctuation in the noise PSD estimation. In the voiced frames where the *a priori* SNR is expected



(a) PESQ (b) FWSegSNR (dB)

(c) SRMR (d) CD (dB)

(d) LLR

Figure 9: Comparison with conventional methods as well as SS-based single-channel methods from the RCSE2014. PESQ values were not reported for all the methods.

to be higher, we floored the gain function at 0.5 which resulted in a positive impact on the FWSegSNR.

We have included two GSS plots with $\alpha = 0.4$ and 0.98 in Fig. 9, however, we shall refer the graph with $\alpha = 0.4$ by the common term GSS in the subsequent discussion, unless specifically mentioned otherwise. From Fig. 9 we observe that GSS shows significant improvements over the degraded speech in terms of all the metrics. It also outperforms conventional SS in most of the cases, especially at the highly reverberant conditions. In comparison with the RCSE2014 methods, GSS consistently performs better in terms of PESQ and CD in highly reverberant conditions, whereas maintaining a comparable performance in terms of other metrics. As the performances of the algorithms fluctuate with different conditions and metrics, we summarize the overall results in Table 4 by taking the average value for each metric and highlighted the best two values with boldfaces. From Table 4 we can observe that GSS outperforms all other methods in terms of PESQ and CD, and remains in the top-two for SRMR and FWSegSNR though shows slight degradation in terms of LLR.

Table 4: Average performance of SS-based methods

Metrics	CD	F. SNR	LLR	PESQ	SRMR
Degraded	3.97	3.62	0.58	1.5	3.68
GSS	**3.54**	**6.62**	0.62	**2.33**	**4.56**
SS	3.56	5.37	**0.56**	**2.23**	4.08
Cauchi	**3.55**	6.09	0.59	1.66	4.29
Veras	4.22	5.16	0.59	-	**4.82**
Wisdom	3.57	**7.07**	0.57	1.69	4.55
Xiao	3.82	4.75	**0.56**	-	4.01

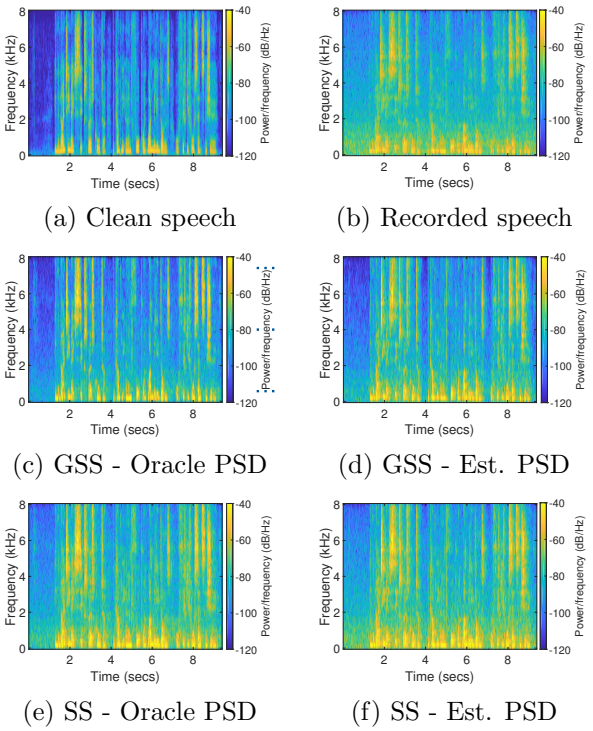
One interesting observation we can make from Fig. 9 that the selection of α directly influences the amount of dereverberation and speech distortion. With a lower α value, speech distortion remains at a lower level (i.e. higher PESQ value) at the cost of small amount of dereverberation (i.e. low SRMR). Hence, α can be controlled to determine the trade off between dereverberation and speech distortion.

The performance issue of GSS in a less reverberant condition can be explained from Table 1 where we observe that the cross-term error is significantly lower at R1 and R2 ($T_{60} = 300$) compared to other room conditions ($T_{60} = 600, 700$). We discussed in Section 3.4 that the relative improvement of GSS is determined by the balance between cross-term improvement and *a priori* estimation error. Hence, in a low reverberant environment where cross-terms remain insignificant, *a priori* estimation error prevails and GSS shows an inferior performance.

Fig. 10 plots the spectrogram of GSS and conventional SS algorithm outputs along with the clean and degraded signals for a random audio stream. We observe that the spectrograms matches the objective evaluation and shows that GSS produces a better spectral map of the clean speech compared to conventional SS. We have also included a few sample audio files for the readers to make subjective evaluations between GSS and conventional SS[1].

Finally, for the completeness of the work, we also include a comparison between GSS and non-SS based single channel algorithms from RCSE2014 [12, 13, 44–48] in Fig. 11. Comparing with the non-SS-based methods, GSS performance is found to be superior compared to the method by Kondo et al. in terms of FWSegSNR, CD, and LLR. Lopez et al. proposed a method which performs similar to GSS in terms of LLR, but GSS performs better in terms of FWSegSNR and CD. The method by Ohtani et al. shows good FWSegSNR, but is affected in terms of CD and LLR which suggests an increased artifacts. Interestingly, Xiao et al. proposed 2 methods

[1] `https://users.cecs.anu.edu.au/~abdullah.fahim/gss/`

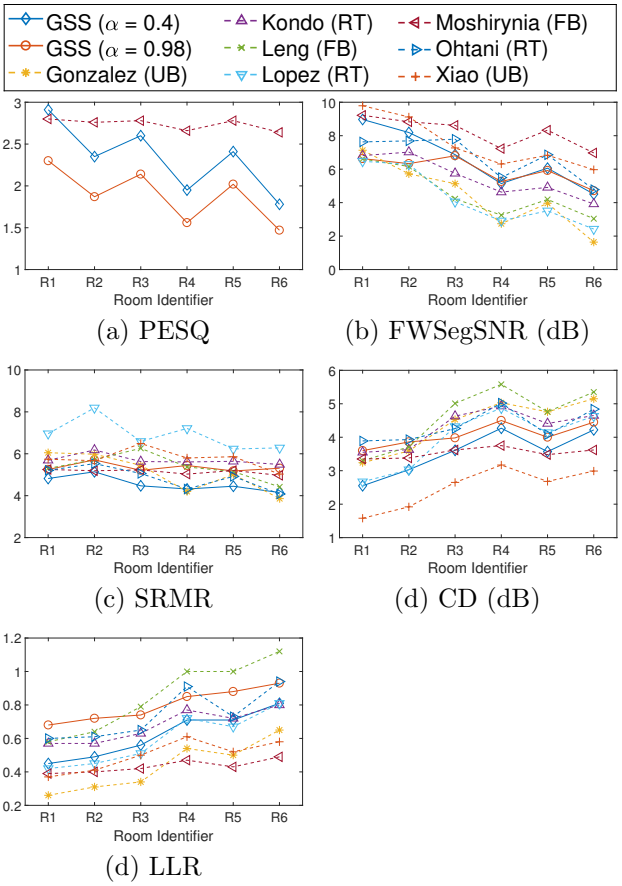


(a) Clean speech (b) Recorded speech

(c) GSS - Oracle PSD (d) GSS - Est. PSD

(e) SS - Oracle PSD (f) SS - Est. PSD

Figure 10: Spectrogram of the clean, degraded, and processed versions of a sample audio.



(a) PESQ (b) FWSegSNR (dB)

(c) SRMR (d) CD (dB)

(d) LLR

Figure 11: Comparison with non-SS based single channel algorithms from the RCSE2014. PESQ values were not reported for all the methods.

in the same paper [12] where the spectral subtraction-based method exhibits inferior performance compared to GSS (Fig. 9 and Table 4) whereas the deep neural network (DNN) based one shows a better performance (Fig. 11). Please note that, the methods proposed by Moshirynia et al. and Leng et al. (Fig 11) used the full batch of audio files to train their model, hence, not comparable with spectral subtraction-based methods which can be designed to work in real time. Note that, Fig. 11 does not necessarily provide a fair comparison due to the differences in the underlying techniques and different resource requirements. In order to compare different classes of algorithms, a detailed analysis of the underlying principles is required which is out of the scope of this work. It is worth mentioning that the main strength of the SS-based methods lies in the simplicity of design, low computational cost and the ability to implement with real-time data.

6 Conclusions

We performed a detailed theoretical analysis and experimental evaluation of GSS with an end-to-end speech enhancement system in realistic noisy and reverberant environments. We determined a fundamental limitation of GSS and explained the importance of *a priori* SNR estimation in that regard. An investigation was carried out on the speech components of noisy reverberant signals to find out the impact of cross-terms in spectral subtraction based techniques.

Through this work, we conducted a series of experiments and concluded that while GSS improves the performance of SS over the conventional methods, it shows a high sensitivity to PSD estimation error. Hence, it was not surprising to find out that GSS outperformed all the conventional SS-based methods when the PSD components were assumed to be known. However, we also observed that, despite its noise sensitivity, GSS performed considerably better compared to other conventional and contemporary SS-based techniques in the the blind dereverberation and noise suppression task of RCSE2014. As the overall performance of GSS depends on the balance between the cross-term improvement and *a priori* estimation error, the improved GSS performances in the RCSE2014 task indicate the significant presence of the cross-term components in most real-world acoustic systems.

References

[1] J. P. A. Lochner, J.F. Burger: The intelligibility of speech under reverberant conditions. Acta Acust united Ac **11** (1961) 195–200.

[2] K. Kinoshita, M. Delcroix, T. Nakatani, M. Miyoshi: Suppression of late reverberation effect on speech sig-

nal using long-term multiple-step linear prediction. IEEE Trans. Audio, Speech, Lang. Process. **17** (2009) 534–545.

[3] T. Arai, N. Hodoshima, K. Yasu: Using steady-state suppression to improve speech intelligibility in reverberant environments for elderly listeners. IEEE Trans. Audio, Speech, Lang. Process. **18** (2010) 1775–1780.

[4] Y. Hu, K. Kokkinakis: Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners. J. Acoust. Soc. of Amer. **135** (2014) EL22–EL28.

[5] M. Miyoshi: Recovering the quality of speech degraded by reverberations in a room. J. Acoust. Soc. of Amer. **120** (2006) 3046–3046.

[6] S. Gannot, M. Moonen: Subspace methods for multimicrophone speech dereverberation. EURASIP J. Appl. Signal Process. **2003** (2003) 1074–1090.

[7] M. Miyoshi, Y. Kaneda: Inverse filtering of room acoustics. IEEE Trans. Acoust., Speech, Signal Process. **36** (1988) 145–152.

[8] B. Schwartz, S. Gannot, E. A. P. Habets: Multimicrophone speech dereverberation using expectation-maximization and Kalman smoothing. Proc. EUSIPCO (2013) 1–5.

[9] T. Yoshioka, T. Nakatani, M. Miyoshi: Integrated speech enhancement method using noise suppression and dereverberation. IEEE Trans. Audio, Speech, Lang. Process. **17** (2009) 231–246.

[10] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, T. Zhang: Learning spectral mapping for speech dereverberation and denoising. IEEE/ACM Trans. Audio, Speech, Lang. Process. **23** (2015) 982–992.

[11] M. Wölfel: Enhanced speech features by single-channel joint compensation of noise and reverberation. IEEE Trans. Audio, Speech, Lang. Process. **17** (2009) 312–323.

[12] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E. Chng, H. Li: The NTU-ADSC systems for reverberation challenge 2014. Proc. REVERB Challenge Workshop, o2.2 (2014).

[13] M. Moshirynia, F. Razzazi, A. Haghbin: A speech dereverberation method using adaptive sparse dictionary learning. Proc. REVERB Challenge Workshop, p1.2 **2** (2014).

[14] T. Nakatani, K. Kinoshita, M. Miyoshi: Harmonicity-based blind dereverberation for single-channel speech signals. IEEE Trans. Audio, Speech, Lang. Process. **15** (2007) 80–95.

[15] M. Delcroix, T. Hikichi, M. Miyoshi: Dereverberation and denoising using multichannel linear prediction. IEEE Trans. Audio, Speech, Lang. Process. **15** (2007) 1791–1801.

[16] M. Delcroix, T. Hikichi, M. Miyoshi: Precise dereverberation using multichannel linear prediction. IEEE Trans. Audio, Speech, Lang. Process. **15** (2007) 430–440.

[17] K. Lebart, J. Boucher, P. N. Denbigh: A new method based on spectral subtraction for speech dereverberation. Acta Acust united Ac **87** (2001) 359–366.

[18] J. Polack: La transmission de l'énergie sonore dans les salles (1988) .

[19] E. A. P. Habets: Single-channel speech dereverberation based on spectral subtraction. Proc. ProRISC (2004) 250–254.

[20] E. Habets, S. Gannot: Dual-microphone speech dereverberation using a reference signal. Proc. IEEE ICASSP **4** (2007) IV–901.

[21] E. Habets, J. Benesty, I. Cohen, S. Gannot, J. Dmochowski: New insights into the MVDR beamformer in room acoustics. IEEE Trans. Audio, Speech, Lang. Process. **18** (2010) 158–170.

[22] E. A. P. Habets: Speech dereverberation using statistical reverberation models. Speech Dereverberation, Springer 2010. pp57–93.

[23] S. Wisdom, T. Powers, L. Atlas, J. Pitton: Enhancement of reverberant and noisy speech by extending its coherence. Proc. REVERB Challenge Workshop, p1.11 (2014).

[24] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, S. Goetze: Joint dereverberation and noise reduction using beamforming and a single-channel speech enhancement scheme. Proc. REVERB Challenge Workshop, o1.2 (2014).

[25] M. Wu, D. Wang: A two-stage algorithm for one-microphone reverberant speech enhancement. IEEE Trans. Audio, Speech, Lang. Process. **14** (2006) 774–784.

[26] J. C. S. Veras, T. D. M. Prego, A. A. D. Lima, T. N. Ferreira, S. L. Netto: Speech quality enhancement based on spectral subtraction. Proc. REVERB Challenge Workshop, p1.7 **7** (2014).

[27] K. Kokkinakis, C. Runge, Q. Tahmina, Y. Hu: Evaluation of a spectral subtraction strategy to suppress reverberant energy in cochlear implant devices. J. Acoust. Soc. of Amer. **138** (2015) 115–124.

[28] Y. Lu, P. C. Loizou: A geometric approach to spectral subtraction. Speech communication **50** (2008) 453–466.

[29] S. V. Vaseghi: Advanced digital signal processing and noise reduction. John Wiley & Sons, 2008. pp297–315.

[30] D. Wang, J. Lim: The unimportance of phase in speech enhancement. IEEE Trans. Audio, Speech, Signal Process. **30** (1982) 679–681.

[31] T. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals: WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition. Proc. IEEE ICASSP **1** (1995) 81–84.

[32] Y. Ephraim, D. Malah: Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Audio, Speech, Signal Process. **32** (1984) 1109–1121.

[33] Y. Ephraim, D. Malah: Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Audio, Speech, Signal Process. **33** (1985) 443–445.

[34] M. Wu, D. Wang: A pitch-based method for the estimation of short reverberation time. Acta Acust united Ac **92** (2006) 337–339.

[35] P. Kendrick, F. F. Li, T. J. Cox, Y. Zhang, J. A. Chambers: Blind estimation of reverberation parameters for non-diffuse rooms. Acta Acust united Ac **93** (2007) 760–770.

[36] H. Löllmann, E. Yilmaz, M. Jeub, P. Vary: An improved algorithm for blind reverberation time estimation. Proc. IWAENC (2010) 1–4.

[37] S. M. Ban, H. S. Kim: Weight-Space Viterbi Decoding Based Spectral Subtraction for Reverberant Speech Recognition. IEEE Signal Process. Lett. **22** (2015) 1424–1428.

[38] E. Verteletskaya, K. Sakhnov: Voice activity detection for speech enhancement applications. Acta Polytechnica **50** (2010).

[39] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, others: A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. EURASIP J. Advances Signal Process. **2016** (2016) 1–19.

[40] ITU-T Rec.: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Rec. ITU-T P. 862. (2001).

[41] T. H. Falk, C. Zheng, W. Chan: A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Trans. Audio, Speech, Lang. Process. **18** (2010) 1766–1774.

[42] E. A. P. Habets: Single and multi-microphone speech dereverberation using spectral enhancement. Dissertation Abstracts International **68** (2007) 127–152.

[43] Y. Hu, P. C. Loizou: Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio, Speech, Lang. Process. **16** (2008) 229–238.

[44] N. Lopez, G. Richard, Y. Grenier, I. Bourmeyster: Reverberation suppression based on sparse linear prediction in noisy environments. Proc. REVERB Challenge Workshop, p2.3 (2014).

[45] K. Ohtani, T. Komatsu, T. Nishino, K. Takeda: Adaptive dereverberation method based on complementary Wiener filter and modulation transfer function. Proc. REVERB Challenge Workshop, p1.5 (2014).

[46] K. Kondo: A computationally restrained and single-channel blind dereverberation method utilizing iterative spectral modifications. Proc. REVERB Challenge Workshop, p2.4 (2014).

[47] Y. R. Leng, J. Dennis, W. Z. T. Ng, T. H. Dat: PBF-GSC Beamforming for ASR and Speech Enhancement in reverberant environments. Proc. REVERB Challenge Workshop, p2.11 (2014).

[48] D. R. González, S. C. Arias, J. R. C. D Lara: Single channel speech enhancement based on zero phase transformation in reverberated environments. Proc. REVERB Challenge Workshop, p2.2 (2014).