SPEECH DEREVERBERATION USING LINEAR PREDICTION WITH ESTIMATION OF EARLY SPEECH SPECTRAL VARIANCE

Mahdi Parchami † *Wei-Ping Zhu*† *Benoit Champagne*⋆

† Department of Electrical and Computer Engineering
Concordia University, Montreal, Quebec, Canada
{m_parch,weiping}@ece.concordia.ca
⋆Department of Electrical and Computer Engineering
McGill University, Montreal, Quebec, Canada
benoit.champagne@mcgill.ca

ABSTRACT

In this paper, we present a new dereverberation algorithm based on the weighted prediction error (WPE) method. In contrast to the conventional WPE method which alternatively estimates the reverberation prediction weights and early speech spectral variance, the proposed algorithm estimates the latter efficiently by employing a geometric spectral enhancement approach and a proper estimate for late reverberant spectral variance (LRSV). Hence, our algorithm does not require iterations to estimate the reverberation prediction weights nor needs alternation between the prediction weights and the spectral variance of early speech. Performance assessments demonstrate considerable improvements in terms of speech quality measures and computational load compared to previous WPE-based dereverberation methods.

Index Terms— Late reverberant spectral variance, linear prediction-based dereverberation, statistical model-based speech enhancement

1. INTRODUCTION

Acoustic signals captured by microphones in an enclosure are often linearly distorted by reflections from walls and other objects. This phenomenon, known as reverberation, degrades the quality and intelligibility of speech, and therefore, restricts the performance of speech processing systems including teleconferencing, voice-controlled systems, hearing aids and automatic speech recognition [1, 2]. Thus, it is essential to mitigate the undesirable effects of reverberation in such applications.

Over the past decades, various dereverberation techniques have been proposed, which can be broadly categorized into acoustic channel equalization, spectral enhancement and probabilistic model-based approaches [3]. There are also a few approaches that emerged recently such as [4, 5]. Although theoretically perfect dereverberation is achievable by acoustic channel equalization, in practice, the performance of such methods is dramatically limited by the accuracy of the estimation of room impulse responses (RIRs) and the need for robust equalization techniques [6]. Further, the spectral enhancement methods originally developed for noise reduction purposes [7], introduce disruptive speech distortion while providing limited reverberation suppression. One of the most favorable methods among the probabilistic model-based approaches is the multi-channel linear prediction (MCLP) which is well suited to the single source noiseless scenario. From a statistical viewpoint, this blind reverberation suppressor is based on maximum likelihood (ML) estimation, assuming an auto-regressive (AR) model for the reverberation process [8, 9]. In its basic short-time Fourier transform (STFT) domain implementation, referred to as the weighted prediction error (WPE) method, an iterative algorithm is used to alternatively estimate the reverberation prediction coefficients and speech spectral variance using batch processing of speech utterances (observations).

Considering the iterative nature of the WPE method, even though it may converge within a small number of iterations, theoretically, there is no guarantee on the convergence of the prediction weights. Moreover, in practice, a speech utterance of at least a few seconds is required for the batch processing step in order to obtain accurate linear prediction (LP) weights. Therefore, the computational burden, increased by the number of iterations, is one of the practical limitations of this method. In this paper, we overcome this limitation by introducing a suitable estimator for the speech spectral variance and integrating it into the WPE method. Specifically, this task is accomplished by resorting to the reverberation suppression within the spectral enhancement literature [7] and employing the statistical model-based estimation of late reverberant spectral variance (LRSV) [10] in order to estimate the speech spectral variance. In addition to the performance merit with respect to the previous WPE-based dereverberation methods, the presented approach offers a considerable gain in reducing the implementation complexity.

2. WPE METHOD

In this section, we give a brief description of the WPE method from a statistical viewpoint, that will help to set the notations for the following sections. Consider a scenario where a single source of speech is captured by microphones, all located within a noiseless reverberant enclosure.. In the STFT domain, we denote the clean speech signal by $s_{n,k}$ with time frame index $n \in \{1, \ldots, N\}$ and frequency bin index $k \in \{1, \ldots, K\}$. Then, the reverberant speech signal observed at the m-th microphone, $x^m_{n,k}$, can be represented in the STFT domain through a linear prediction model as [9]

$$x^m_{n,k} = \sum_{l=0}^{L_h\text{-}1} \left(h^m_{l,k}\right)^* s_{n-l,k} + e^m_{n,k}, \tag{1}$$

where $h^m_{l,k}$ is an approximation of the acoustic transfer function (ATF) between the speech source and the m-th microphone in the STFT domain with length of L_h, and $(.)^*$ denotes the complex conjugate operator. The additive term $e^m_{n,k}$ models the linear prediction error and is often disregarded in the STFT domain [9, 11]. There-

fore, the model in (1) can be rewritten as

$$x^m_{n,k} = d^m_{n,k} + \sum_{l=D}^{L_h\text{-}l} \left(h^m_{l,k}\right)^* s_{n-l,k}, \tag{2}$$

where $d^m_{n,k} = \sum_{l=0}^{D-1} \left(h^m_{l,k}\right)^* s_{n-l,k}$ is the sum of anechoic (direct) speech and early reflections at the m-th microphone, and D corresponds to the duration of the early reflections. Most dereverberation techniques, including the WPE method, aim at reconstructing $d_{n,k}$ as the desired signal, since the early reflections actually improve speech intelligibility and also the SNR in noisy environments [12]. Replacing the convolutive model in (2) by an auto-regressive model results in the well-known multi-channel linear prediction (MCLP) form for the observation at the first microphone, as the following [9]

$$d^1_{n,k} = x^1_{n,k} - \sum_{m=1}^{M} \left(\mathbf{g}^m_k\right)^{\mathrm{H}} \mathbf{x}^m_{n-D,k}, \tag{3}$$

where $d^1_{n,k} \equiv d_{n,k}$ is the desired signal, $(.)^{\mathrm{H}}$ is the hermitian operator, and the vectors $\mathbf{x}^m_{n-D,k}$ and $\mathbf{g}^m_k$ are defined as

$$\begin{aligned} \mathbf{x}^m_{n-D,k} &= [x^m_{n-D,k}\ ,\ x^m_{n-D-1,k}\ ,\ \ldots,\ x^m_{n-D-(L_k-1),k}]^{\mathrm{T}}, \\ \mathbf{g}^m_k &= [g^m_{0,k}\ ,\ g^m_{1,k}\ ,\ \ldots,\ g^m_{L_k-1,k}]^{\mathrm{T}}. \end{aligned} \tag{4}$$

$\mathbf{g}^m_k$ is the regression vector (reverberation prediction weights) of order L_k for the m-th channel. Considering each of the vector sets $\{\mathbf{x}^m_{n-D,k}\}$ and $\{\mathbf{g}^m_k\}$ for $m = 1, 2, \ldots, M$ and concatenating them over m to respectively form $\mathbf{x}_{n-D,k}$ and $\mathbf{g}_k$, (3) is written in the compact form

$$d_{n,k} = x^1_{n,k} - \mathbf{g}^{\mathrm{H}}_k\, \mathbf{x}_{n-D,k}, \tag{5}$$

Estimation of the regression vector $\mathbf{g}_k$ and using it in (5) gives the WPE estimate of the desired speech. From a statistical viewpoint, this is performed in [9] by using the maximum likelihood (ML) estimation of the desired speech $d_{n,k}$ at each frequency bin. The conventional WPE method [8, 9] assumes a circular complex Gaussian distribution for each of the desired speech STFT coefficients, $d_{n,k}$, with time-varying power spectrum and zero mean. Assuming independence across time frames, n, the joint distribution of the desired speech coefficients at frequency bin, k, is given by

$$p(\mathbf{d}_k) = \prod_{n=1}^{N} p(d_{n,k}) = \prod_{n=1}^{N} \frac{1}{\pi\sigma^2_{d_{n,k}}} \exp\left(-\frac{|d_{n,k}|^2}{\sigma^2_{d_{n,k}}}\right), \tag{6}$$

with $\sigma^2_{d_{n,k}}$ being the time-varying spectral variance of the desired speech. Now, by inserting $d_{n,k}$ from (5) into (6), it can be observed that the set of unknown parameters at each frequency bin consists of the regression vector, $\mathbf{g}_k$, and the desired speech spectral variances $\sigma^2_{\mathbf{d}_k} = \{\sigma^2_{d_{1,k}}, \sigma^2_{d_{2,k}}, \ldots, \sigma^2_{d_{N,k}}\}$. Denoting this set by $\Theta_k = \{\mathbf{g}_k, \sigma^2_{\mathbf{d}_k}\}$, and taking the negative of logarithm of $p(\mathbf{d}_k)$ in (6), the objective function for the parameter set Θ_k can be written as

$$\begin{aligned} \mathcal{J}(\Theta_k) = -\log\ p(\mathbf{d}_k|\Theta_k) = \\ \sum_{n=1}^{N} \left(\log\ \sigma^2_{d_{n,k}} + \frac{\left|x^1_{n,k} - \mathbf{g}^{\mathrm{H}}_k\, \mathbf{x}_{n-D,k}\right|^2}{\sigma^2_{d_{n,k}}}\right). \end{aligned} \tag{7}$$

where we neglected the constant terms. Here, (7) has to be minimized at each frequency bin with respect to the parameter set Θ_k. Even though minimization of (7) over $\sigma^2_{\mathbf{d}_k}$ or $\mathbf{g}_k$ is straightforward, a joint optimization of (7) with respect to both subsets of parameters cannot be performed analytically. Thus, a two-step process is used in [8, 9] wherein one of the two parameter subsets is estimated at each step by means of an ML estimation, given an estimate of the other. The entire process is iteratively continued until some convergence criterion is satisfied or a maximum number of iterations is reached. More specifically, at the i-th iteration, (7) is minimized with respect to the set of variances $\sigma^2_{\mathbf{d}_k}$, leading to the following simple estimate for each $\sigma^2_{d_{n,k}}$

$${\sigma^2_{d_{n,k}}}^{(i)} = \left|{d_{n,k}}^{(i-1)}\right|^2, \quad n = 1, 2, \ldots, N. \tag{8}$$

where ${d_{n,k}}^{(i-1)}$ is the desired speech estimate from the previous iteration. For the first iteration, it is suggested in [8, 9] to use $x^1_{n,k}$ as the initial value for this term. Next, minimization of (7) with respect to the regression vector $\mathbf{g}_k$ results in a conventional least squares problem with the following closed-form solution,

$${\mathbf{g}_k}^{(i)} = \left(\sum_{n=1}^{N} \frac{\mathbf{x}_{n-D,k}\mathbf{x}^{\mathrm{H}}_{n-D,k}}{{\sigma^2_{d_{n,k}}}^{(i)}}\right)^{-1} \sum_{n=1}^{N} \frac{\mathbf{x}_{n-D,k}\left(x^1_{n,k}\right)^*}{{\sigma^2_{d_{n,k}}}^{(i)}}, \tag{9}$$

The estimated $\mathbf{g}_k$ above is then used in (5) to obtain $d_{n,k}$'s, which in turn, are exploited to estimate $\sigma^2_{d_{n,k}}$'s for the next iteration, according to (8). Often in practice, 3 to 5 iterations lead to the best possible results [12], yet, there is no guarantee or a widely accepted criterion on the convergence of the method. Applying more iterations does not necessarily result in improvements and may even degrade the performance. Furthermore, instantaneous estimates of the desired speech variance as given by (8) may lead to very small values that deteriorate the overall performance. The aforementioned disadvantages can be mitigated by employing a proper estimate of the spectral variance of desired speech, as explained in the following section.

3. WPE BASED ON EARLY REVERBERANT SPECTRAL VARIANCE ESTIMATION

In this section, we propose an efficient estimator for the spectral variance of the desired speech, $\sigma^2_{d_{n,k}}$, based on the statistical modeling of ATF, and incorporate this estimator within the WPE dereverberation algorithm. As seen from (1)-(2), the desired speech $d_{n,k}$ is in fact the sum of the first D delayed and weighted clean speech terms, $s_{n-l,k}$. In the context of statistical spectral enhancement methods [7, 10], $d_{n,k}$ is often referred to as early speech, as compared to late reverberant speech given by the sums in (2) and (3). Therefore, the observation at the first microphone can be rewritten as

$$x^1_{n,k} = d_{n,k} + r_{n,k}, \tag{10}$$

with $r_{n,k}$ denoting the late reverberant speech. Several methods are available in the spectral enhancement literature for the estimation of $\sigma^2_{d_{n,k}}$ in (10), such as the decision directed (DD) approach for direct-to-reverberant ratio (DRR) estimation [7]. Using this method, $\sigma^2_{d_{n,k}}$ can be obtained as the product of the estimated DRR, i.e., $\sigma^2_{d_{n,k}}/\sigma^2_{r_{n,k}}$, and an estimate of the late reverberant spectral variance, $\sigma^2_{r_{n,k}}$. However, the application of conventional spectral enhancement techniques, originally developed for noise reduction purposes, is based on the assumption of independence between $d_{n,k}$ and $r_{n,k}$. Here, however, contrary to the scenario of additive noise, as evidenced from the model in (1) and (2), the early and late reverberant terms are basically correlated, due to the temporal correlation across successive time frames of speech signal. Therefore, the non-zero correlation between $d_{n,k}$ and $r_{n,k}$ must be taken into account. Doing so, it follows from (10) that

$$\sigma^2_{x^1_{n,k}} = \sigma^2_{d_{n,k}} + \sigma^2_{r_{n,k}} + 2E\{\Re\{d_{n,k}\, r^*_{n,k}\}\}, \tag{11}$$

with $\Re\{.\}$ denoting the real value and $2E\{\Re\{d_{n,k}\ r^*_{n,k}\}\}$ representing the non-zero cross-correlation terms between $d_{n,k}$ and $r_{n,k}$. Nevertheless, the estimation of the cross-correlation terms in (11), due to their dependency on the phases of $d_{n,k}$ and $r_{n,k}$, may not be analytically tractable.

In [13], a spectral subtraction algorithm for noise suppression has been proposed based on the deterministic estimation of speech magnitudes in terms of observation and noise magnitudes without assuming that they are independent. Therein, the authors consider the following similar problem to (11), that is

$$\left|x^1_{n,k}\right|^2 = |d_{n,k}|^2 + |r_{n,k}|^2 + 2\,|d_{n,k}|\,|r_{n,k}|\cos\left(\theta_{d_{n,k}} - \theta_{r_{n,k}}\right). \quad (12)$$

where $|d_{n,k}|$ is to be estimated in terms of $|x^1_{n,k}|$ and $|r_{n,k}|$, and $\theta_{d_{n,k}}$ and $\theta_{r_{n,k}}$ are the unknown phases of $d_{n,k}$ and $r_{n,k}$ respectively. Through a geometric approach, the following estimate of $|d_{n,k}|$ is then obtained

$$\left|\hat{d}_{n,k}\right| = \sqrt{\frac{1 - \frac{(\gamma-\xi+1)^2}{4\gamma}}{1 - \frac{(\gamma-\xi-1)^2}{4\xi}}}\ \left|x^1_{n,k}\right|, \quad (13)$$

where the two parameters ξ and γ are defined as

$$\xi_{n,k} \triangleq \frac{|d_{n,k}|^2}{|r_{n,k}|^2}\ ,\ \gamma_{n,k} \triangleq \frac{\left|x^1_{n,k}\right|^2}{|r_{n,k}|^2}, \quad (14)$$

Herein, we propose to employ this approach in order to provide a correlation-aware estimate of $|d_{n,k}|$, to be exploited in turn in the estimation of $\sigma^2_{d_{n,k}}$.

Due to the unavailability of $|d_{n,k}|^2$ and $|r_{n,k}|^2$, the two parameters in (14) are not available *a priori* and have to be substituted by their approximations. To this end, we exploit $|\hat{d}_{n-1,k}|^2$ for $|d_{n,k}|^2$ and a short-term estimate of $\sigma^2_{r_{n,k}}$ for $|r_{n,k}|^2$. To determine the latter, we resort to the statistical model-based estimation of the LRSV, which has been widely used in the spectral enhancement literature. Therein, an estimate of this key parameter is derived using a statistical model for the ATF along with recursive smoothing schemes. In brief, the following scheme is conventionally used to estimate the LRSV [10]:

$$\sigma^2_{x^1_{n,k}} = (1-\beta)\ \sigma^2_{x^1_{n-1,k}} + \beta\left|x^1_{n,k}\right|^2, \quad (15a)$$

$$\sigma^2_{\tilde{r}_{n,k}} = (1-\kappa)\ \sigma^2_{\tilde{r}_{n-1,k}} + \kappa\,\sigma^2_{x^1_{n-1,k}}, \quad (15b)$$

$$\sigma^2_{r_{n,k}} = e^{-2\alpha_k R N_e}\ \sigma^2_{\tilde{r}_{n-(N_e-1),k}}. \quad (15c)$$

where α_k is related to the 60 dB reverberation time, $T_{60dB,k}$, through $\alpha_k = 3\log 10/\left(T_{60dB,k} f_s\right)$ with f_s as the sampling frequency in Hz, R is the STFT time shift in samples, β and κ are smoothing parameters (which can be in general frequency-dependent) and N_e is the delay parameter defining the number of assumed early speech frames, which is herein taken as D. The term $\tilde{r}_{n,k}$ actually represents the entire reverberant speech including both early and late reverberant terms, but excluding the first early term. Using the LRSV estimator in (15), the short-term estimate of $\sigma^2_{r_{n,k}}$ is obtained by choosing the smoothing parameters β and κ close to one. In this way, the estimate of $\sigma^2_{r_{n,k}}$ is updated faster, and is therefore closer to the true value of $|r_{n,k}|^2$. Yet, to avoid too small values of the approximated $|r_{n,k}|^2$ in the denominator of (14), a lower bound is applied to this quantity.

Now, given the estimate of early speech magnitude, $|\hat{d}_{n,k}|$, provided by (13), it is simple to use a recursive smoothing scheme to estimate $\sigma^2_{d_{n,k}}$, as the following

$$\hat{\sigma}^2_{d_{n,k}} = (1-\eta)\,\hat{\sigma}^2_{d_{n-1,k}} + \eta\,|\hat{d}_{n,k}|^2, \quad (16)$$

with η as a fixed smoothing parameter. This estimate of $\sigma^2_{d_{n,k}}$ can be efficiently integrated into the WPE method discussed in Section 2, replacing the instantaneous estimate given by (8). By doing so, the objective function in (7) turns into a function of only the regression vector, $\mathbf{g}_k$, and it is possible to obtain the latter without iterations by (9).

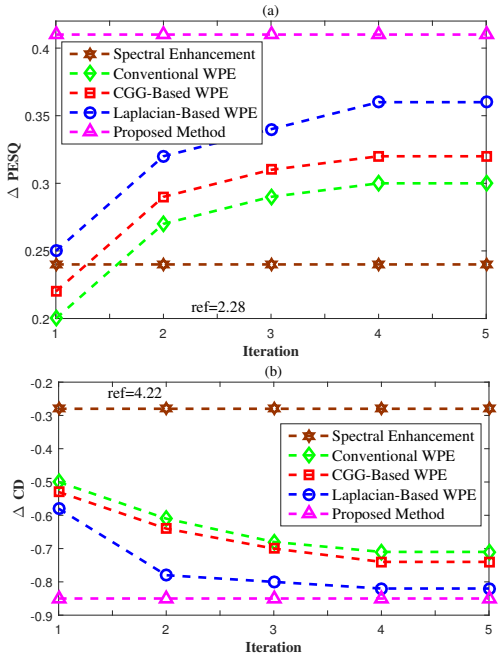

Fig. 1: Improvement in (a): PESQ and (b): CD measures versus the number of iterations for different methods. The reported values are obtained as the averaged improvements over all utterances.

4. PERFORMANCE EVALUATION

To evaluate the performance of the proposed approach, clean speech utterances were used from the TIMIT database [14], including 10 male and 10 female speakers. The sampling rate was set to 16 kHz and a 50 ms Hamming window with the overlap of 75% was used for the STFT analysis-synthesis. The number of early speech terms, i.e. D, was set to 3 and to obtain the best achievable performance, we used the first 10 second segment of the reverberant speech to estimate $\mathbf{g}_k$. To implement the proposed method in Section 3, we set the minimum value of the estimated $\sigma^2_{r_{n,k}}$ from (11) to $\epsilon = 10^{-3}$ before using it in (15). As for the recursive smoothing schemes in (15), we took β and κ to be 0.50 and 0.80, respectively, and chose N_e as 3. Also, the reverberation time, $T_{60dB,k}$, in (15) was estimated blindly by the approach introduced in [15]. Our method requires no prior knowledge of the DRR parameter.

To generate reverberant noisy speech signals for the scenario in Fig. 1, we convolved the clean speech utterances with measured RIRs and added real-world noise with SNR of 10dB. The RIRs and recorded noise were taken from the SimData of the REVERB challenge [16], where an 8 channel circular array with diameter of 20cm

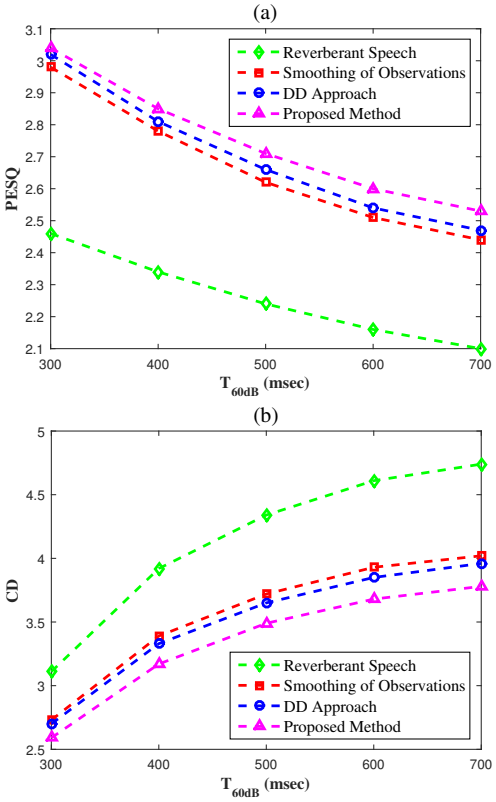

Fig. 2: (a): PESQ and (b): CD measures versus T_{60dB} for the reverberant speech and the enhanced one using the WPE method with different estimators of the early speech spectral variance, $\sigma^2_{d_{n,k}}$.

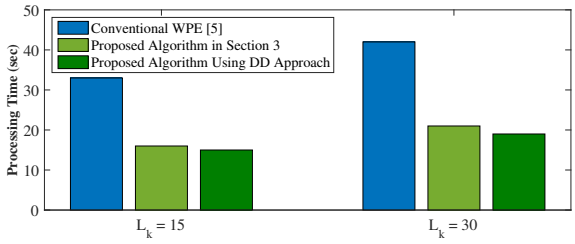

Fig. 3: Processing time required for the estimation of $\mathbf{g}_k$ with lengths of $L_k = 15$ and $L_k = 30$ from a 10 secs. speech segment for different methods. A i5-2400 CPU @ 3.10GHz with RAM: 4.00GB was used for the implementation using Matlab.

was used in a 3.7m×5.5m room. In Fig. 1, performance comparison of the proposed method with respect to the conventional, two recent WPE-based methods, and using spectral enhancement [7] is illustrated in terms of PESQ (perceptual evaluation of speech quality) and CD (cepstral distance) [17]. The values of ΔPESQ and ΔCD represent the improvements in these quantities relative to the corresponding value for the reverberant speech, denoted in the figure as "ref". Recently, there has been growing interest in the use of properly fitted distributions for speech priors and estimation of their parameters. In the context of WPE-based methods, this has been accomplished by using Laplacian and complex generalized Gaussian (CGG) priors, respectively in [18] and [19]. As seen in Fig. 1, the Laplacian-based method is capable of providing considerable performance improvements with respect to the conventional WPE in Section 2, whereas the CGG-based gives trivial improvements. The latter, as concluded in [19], is actually similar to the same conventional WPE method but with a different instantaneous estimator for the desired speech spectral variance, $\sigma^2_{d_{n,k}}$. However, the proposed method in Section 3, in addition to being non-iterative, is able to provide a more efficient and accurate estimate of $\sigma^2_{d_{n,k}}$. It is also observable that the spectral enhancement method with an LRSV estimator is not as efficient as the WPE-based methods for the purpose of dereverberation.

Next, to evaluate experimentally the efficiency of the proposed estimate for $\sigma^2_{d_{n,k}}$ in Section 3, we considered two other recursive smoothing-based schemes to update $\sigma^2_{d_{n,k}}$ and compared their performance with the proposed one in Fig. 2. In this case, to generate reverberant speech signals with controllable amount of reverberation, we convolved synthetic RIRs generated by the image source method (ISM) [20] with the anechoic speech utterances. We considered a scenario where the speech source is 1.5 m away from the two omni-directional microphones capturing the reverberant speech. As in [7], we used the well-known DD approach to estimate the ratio $\sigma^2_{d_{n,k}}/\sigma^2_{r_{n,k}}$ and then multiplied it by the LRSV estimate from (11) to obtain an estimate of $\sigma^2_{d_{n,k}}$, which is denoted in Fig. 2 as the "DD Approach". To demonstrate the importance of taking into account the cross-correlation terms between the desired and late reverberant speech, as in (12), we estimated the desired spectral variance, $\sigma^2_{d_{n,k}}$, by disregarding the cross terms in (12) and using $\sigma^2_{x^1_{n,k}} - \hat{\sigma}^2_{r_{n,k}}$. Since the observation spectral variance is estimated by a fixed recursive smoothing scheme, we denoted this method by "Smoothing of Observations". Referring to Fig. 2, it is observable that the proposed estimation of $\sigma^2_{d_{n,k}}$ results in further reverberation suppression, especially for higher reverberation conditions where the amount of correlation between the desired and late reverberant speech signals increases. It should be noted that the proposed estimator of the desired speech spectral variance can also be used in spectral enhancement-based methods, yet, the dereverberation performance of the latter was found to be inferior to the LP-based methods. We also evaluated experimentally the computational cost of our proposed algorithm using the estimation of $\sigma^2_{d_{n,k}}$ discussed in Section 3, proposed algorithm using the DD approach to estimate $\sigma^2_{d_{n,k}}$; and the conventional WPE method using a maximum of 3 iterations. The results are presented in Fig. 3 in terms of the batch processing time needed to estimate the WPE regression vector. As seen, by eliminating the iterative process of the WPE method through the proposed algorithm, the computational effort has been considerably reduced.

5. CONCLUSION

We presented a novel LP-based dereverberation algorithm by proposing an efficient estimation scheme for the spectral variance of early speech. The spectral variance estimate is obtained through a geometric spectral enhancement approach and a conventional LRSV estimator, based on the correlation between the early and late reverberant terms. Using the proposed algorithm, the well-known WPE method can be implemented in a non-iterative fashion, and thanks to the efficient estimate of the early speech spectral variance, significant improvements in speech quality and reduction in implementation complexity can be achieved with respect to the previously proposed WPE-based methods.

6. REFERENCES

[1] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, Nov 2012.

[2] R. Maas, E. A. P. Habets, A. Sehr, and W. Kellermann, "On the application of reverberation suppression to robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 297–300.

[3] P.A. Naylor and N.D. Gaubitch, Eds., *Speech Dereverberation*, Springer-Verlag, London, 2010.

[4] D. Schmid, S. Malik, and G. Enzner, "An expectation-maximization algorithm for multichannel adaptive speech dereverberation in the frequency-domain," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 17–20.

[5] M. Togami and Y. Kawaguchi, "Noise robust speech dereverberation with Kalman smoother," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 7447–7451.

[6] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1879–1890, Sept 2013.

[7] E.A.P. Habets, *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*, Ph.D. thesis, Technische Universiteit Eindhoven, Netherlands, 2007.

[8] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and Biing-Hwang Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, March 2008, pp. 85–88.

[9] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and Biing-Hwang Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, Sept 2010.

[10] E.A.P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, Sept 2009.

[11] Y. Iwata and T. Nakatani, "Introduction of speech log-spectral priors into dereverberation based on itakura-saito distance minimization," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, March 2012, pp. 245–248.

[12] I. Cohen, J. Benesty, and S. Gannot, Eds., *Speech Processing in Modern Communication*, chapter 6, Springer-Verlag, Berlin Heidelberg, 2010.

[13] Y. Lu and P.C. Loizou, "A geometric approach to spectral subtraction," *Speech Communication*, vol. 50, no. 6, pp. 453–466, 2008.

[14] J.S. Garofolo et al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1," Philadelphia: Linguistic Data Consortium, 1993.

[15] H.W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Aug 2010.

[16] "SimData: dev and eval sets based on WSJCAM0", *REVERB Challenge*, 2013 (accessed January 12, 2016), available at http://reverb2014.dereverberation.com/download.html.

[17] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The reverb challenge: Acommon evaluation framework for dereverberation and recognition of reverberant speech," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2013, pp. 1–4.

[18] A. Jukic and S. Doclo, "Speech dereverberation using weighted prediction error with laplacian model of the desired signal," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, May 2014, pp. 5172–5176.

[19] A. Jukic, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1509–1520, Sept 2015.

[20] E.A. Lehmann, "Image-source method: Matlab code implementation," available at http://www.eric-lehmann.com/.