Speech Communication 87 (2017) 49–57

Contents lists available at ScienceDirect

Speech Communication

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

Speech dereverberation using weighted prediction error with correlated inter-frame speech components

Mahdi Parchami[a,*], Wei-Ping Zhu[a], Benoit Champagne[b]

[a] Department of Electrical and Computer Engineering, Concordia University, Montreal, Quebec, Canada
[b] Department of Electrical and Computer Engineering, McGill University, Montreal, Quebec, Canada

ARTICLE INFO

Article history:
Received 18 June 2016
Revised 11 November 2016
Accepted 5 January 2017
Available online 6 January 2017

Keywords:
Inter-frame correlation
Multi-channel linear prediction (MCLP)
Speech dereverberation
Speech enhancement

ABSTRACT

In this paper, we propose a new dereverberation approach based on the weighted prediction error (WPE) method implemented in the short-time Fourier transform (STFT) domain. Our main contribution is to model the temporal correlation of the STFT coefficients across analysis frames, referred to as inter-frame correlation (IFC), and exploit it in the dereverberation process. Since accurate modeling of the IFC is not tractable, we consider an approximate model wherein only a finite number of consecutive speech frames are considered correlated. It is shown that, given an estimate of the IFC matrix, the proposed approach results in a convex quadratic optimization problem with respect to the reverberation prediction weights, and a closed-form solution can be accordingly derived. Furthermore, an efficient method for the estimation of the underlying IFC matrix is developed based on the extension of a recently proposed speech variance estimator. We evaluate the performance of our approach incorporating the estimated IFC matrix and compare it to the original and several variants of the WPE method. The results reveal lower residual reverberation and higher overall quality of the enhanced speech when the proposed method is employed.

1. Introduction

Speech signals captured by distant microphones in an acoustic environment are subject to reflections from the surrounding surfaces and objects, such as walls or furniture. The consequence of this phenomenon, known as reverberation, can deteriorate the perceived quality/intelligibility of the captured speech and also can degrade to a large extent the performance of speech communication systems such as hearing aids, hands-free teleconferencing, source separation, passive source localization, and automatic speech recognition systems (Naylor and Gaubitch, 2010; Yoshioka et al., 2012). Therefore, to improve the overall quality of speech signals in these important applications, highly efficient reverberation suppression algorithms are required.

Over the past three decades, various single- and multi-microphone dereverberation techniques have been developed, which can be broadly classified into: blind system identification and inversion, multi-channel spatial processing, spectral enhancement and probabilistic model-based approaches (Naylor and Gaubitch, 2010; Habets, 2007). In addition, a few alternative dereverberation approaches emerged recently, such as those based on the use of expectation-maximization (EM) and Kalman filtering algorithms, as reported in Schmid et al. (2012) and Togami and Kawaguchi (2013). Among all these techniques, the model-based statistical approaches, which seek to optimally estimate the anechoic speech, have attracted considerable interest as further discussed below.

In Attias et al. (2001), probabilistic models of speech were incorporated into a variational Bayesian EM algorithm which estimates the source signal, the acoustic channel and all the involved parameters in an iterative manner. A different strategy was followed in Yoshioka et al. (2009), where the parameters of all-pole models for both the desired speech signal and the reverberation component are iteratively determined by maximizing the likelihood function of the model parameters through an EM approach. In this way, a minimum mean-squared error (MMSE) estimator is

Abbreviations: ATF, acoustic transfer function; AR, auto-regressive; CD, cepstrum distance; CGG, complex generalized Gaussian; DRR, direct to reverberant ratio; EM, expectation-maximization; FW-SNR, frequency-weighted segmental SNR; IFC, inter-frame correlation; ISM, image source method; LRSV, late reverberation spectral variance; LPC, linear prediction coefficients; ML, maximum likelihood; MMSE, minimum mean-squared error; MCLP, multi-channel linear prediction; PESQ, perceptual evaluation of speech quality; RIR, room impulse response; STFT, short-time Fourier transform; SRMR, signal to reverberation modulation energy ratio; SNR, signal to noise ratio; WPE, weighted prediction error.

* Corresponding author.
E-mail addresses: m_parch@ece.concordia.ca (M. Parchami), weiping@ece.concordia.ca (W.-P. Zhu), benoit.champagne@mcgill.ca (B. Champagne).

http://dx.doi.org/10.1016/j.specom.2017.01.004
0167-6393/

derived that yields the enhanced speech. In a similar way, by using a time-varying statistical model for the speech and a multi-channel linear prediction (MCLP) model for the reverberation, an efficient dereverberation approach has been developed in Nakatani et al. (2008a) and Kinoshita et al. (2009). Since the implementation of such methods in the time domain is computationally expensive, it was proposed in Nakatani et al. (2008b, 2010) to employ the MCLP-based method in the short-time Fourier transform (STFT) domain. The resulting approach, referred to as the weighted prediction error (WPE) method, is an iterative algorithm that alternatively estimates the reverberation prediction coefficients and speech spectral variance, using batch processing of the speech utterance.

Basically, the WPE method and its variants consider temporally/spectrally independent speech components in the STFT domain. This assumption, despite greatly simplifying the derivation and application of the WPE method, is inaccurate and lacks the modeling of inherent dependencies across time frames and spectral components at each time frame. In Erkelens and Heusdens (2010), it was shown that the STFT coefficients of anechoic speech exhibit significant correlation in time, even with frame overlaps of less than 50%. This correlation, referred to as inter-frame correlation (IFC), is further pronounced in case of highly reverberant speech, due to the convolutive nature of the reverberation. In Habets et al. (2012), in the context of multi-microphone noise reduction in a reverberant environment, it was demonstrated that the achievable performance in terms of noise reduction and speech distortion can be further improved by exploiting IFC. The noise reduction problem using IFC has also been addressed partially in Esch (2012) where, in the propagation step of a noise reduction method based on Kalman filter, the complex-valued prediction weight is used to exploit the temporal correlation of successive speech and noise STFT coefficients. However, similar to Habets et al. (2012), this work assumes perfect knowledge of the theoretical IFC in the derivation of various enhancement algorithms. In summary, the IFC has not been fully explored in the context of STFT domain speech enhancement and the accurate modeling and applications of the speech IFC remains an attractive area for future research, especially in the context of dereverberation where the channel impulse responses are characterized by long memory (Vaseghi, 2006).

In this work, in order to take into account the considerable IFC present in the desired speech (due to the speech characteristics, STFT framing overlaps and heavy reverberation), we reformulate the WPE method through the introduction of an approximate model for the joint probability distribution of the desired speech STFT coefficients within finite segments, each consisting of consecutive frames. Following an ML approach similar to the original WPE method, it is shown that the resulting dereverberation problem leads to a convex optimization problem with a closed-form solution for the reverberation prediction weights, since it can be solved efficiently in a single attempt, unlike the original WPE method whose solution requires an iterative procedure. In addition, regarding the estimation of the underlying IFC matrix for the desired speech component, an extension of the method for speech spectral variance estimation in Parchami et al. (2016) is proposed. The proposed method can efficiently eliminate the reverberant component from the observed speech, prior to the estimation of the cross-spectral variance of the desired speech, that is performed by a first order smoothing scheme. Finally, we evaluate the performance of our approach incorporating the estimated IFC matrix and compare it to the original and several variants of the WPE method. The results reveal lower residual reverberation and higher overall quality of the enhanced speech when the proposed method is employed.

The remainder of this paper is organized as follows. In Section 2, a brief overview of the WPE method is presented. In Section 3, a closed-form solution for the optimum reverberation prediction weights in the WPE method with IFC is developed and a novel technique for the estimation of the IFC matrix is presented. The objective performance evaluation of the proposed approach using different types of reverberant speech signals is discussed in Section 4. Finally, a brief conclusion is given in Section 5.

2. A brief review of the WPE method

Suppose that a speech signal emanating from a single source is captured by M microphones located in a reverberant enclosure. In the STFT domain, we denote the clean speech signal by $s_{n,k}$ with time frame index $n \in \{1,\dots,N\}$ and frequency bin index $k \in \{1,\dots,K\}$ where N is the total number of frames and K is the number of available frequency bins. Then, the reverberant speech signal observed at the mth microphone, $x_{n,k}^{(m)}$, can be represented in the STFT domain using a linear prediction model as Nakatani et al. (2010)

$$x_{n,k}^{(m)} = \sum_{l=0}^{L_h-1} h_{l,k}^{(m)*} s_{n-l,k} + e_{n,k}^{(m)} \tag{1}$$

where $h_{l,k}^{(m)}$ is an approximation of the acoustic transfer function (ATF) between the speech source and the mth microphone in the STFT domain, L_h denotes the length of the ATF (measured in frames) and * denotes the complex conjugate. The additive term $e_{n,k}^{(m)}$ models the linear prediction error and the additive noise and is neglected here as in Nakatani et al. (2010). Therefore, (1) can be rewritten as

$$x_{n,k}^{(m)} = d_{n,k}^{(m)} + \sum_{l=D}^{L_h-1} h_{l,k}^{(m)*} s_{n-l,k} \tag{2}$$

where $d_{n,k}^{(m)} = \sum_{l=0}^{D-1} h_{l,k}^{(m)*} s_{n-l,k}$ is the sum of anechoic (direct-path) speech and early reflections at the mth microphone and D corresponds to the duration of the early reflections. Most dereverberation techniques, including the WPE method, aim at reconstructing the desired signal, say $d_{n,k} \equiv d_{n,k}^{(1)}$, or suppressing the late reverberant terms represented by the summation in (2). Replacing the convolutive model in (2) by an auto-regressive (AR) model results in the well-known multi-channel linear prediction (MCLP) form for the observation at the first microphone, i.e.,

$$d_{n,k} = x_{n,k}^{(1)} - \sum_{m=1}^{M} \mathbf{g}_k^{(m)H} \mathbf{x}_{n,k}^{(m)} = x_{n,k}^{(1)} - \mathbf{G}_k^H \mathbf{X}_{n,k} \tag{3}$$

with superscript H as the Hermitian transpose and the vectors $\mathbf{x}_{n,k}^{(m)}$ and $\mathbf{g}_k^{(m)}$ are defined as

$$\begin{aligned} \mathbf{g}_k^{(m)} &= [g_{0,k}^{(m)}\ ,\ g_{1,k}^{(m)}\ ,\ \dots,\ g_{L_k-1,k}^{(m)}]^T \\ \mathbf{x}_{n,k}^{(m)} &= [x_{n-D,k}^{(m)}\ ,\ x_{n-D-1,k}^{(m)}\ ,\ \dots,\ x_{n-D-(L_k-1),k}^{(m)}]^T \end{aligned} \tag{4}$$

where $\mathbf{g}_k^{(m)}$ is the regression vector (reverberation prediction weights) of order L_k for the mth channel and the superscript T denotes transpose. The right-hand side of (3) has been obtained by concatenating $\{\mathbf{x}_{n,k}^{(m)}\}$ and $\{\mathbf{g}_k^{(m)}\}$ over m to respectively form $\mathbf{X}_{n,k}$ and $\mathbf{G}_k$. Estimation of the regression vector $\mathbf{G}_k$ and insertion of it in (3) can provide an estimate of the desired (dereverberated) speech. From a statistical viewpoint, estimation of $\mathbf{G}_k$ can be performed by applying the maximum likelihood (ML) criterion at each frequency bin. To this end, the conventional WPE method (Nakatani et al., 2008b, 2010) assumes a circular complex Gaussian distribution for the desired speech coefficients, $d_{n,k}$, with (unknown) time-varying spectral variance $\sigma_{d_{n,k}}^2 = E\{|d_{n,k}|^2\}$ and zero mean. Assuming that the desired speech STFT coefficients $d_{n,k}$ are independent across frames, i.e., using zero IFC, the joint distribution of the desired

Table 1
Outline of the steps of the conventional WPE method with temporally independent speech STFT coefficients.

- At each frequency bin k, consider the speech observations $x_{n,k}^{(m)}$, for all n and m and the parameters D, L_k and ϵ.
- Initialize $\sigma_{d_{n,k}}^2$ by $\sigma_{d_{n,k}}^{2[1]} = |x_{n,k}|^2$.
- For, $j = 1, 2, \cdots, J$ (with a fixed number of iterations, J) , repeat the following:
  $\mathbf{A}_k^{[j]} = \sum_{n=1}^{N} \sigma_{d_{n,k}}^{-2[j]} \mathbf{X}_{n,k} \mathbf{X}_{n,k}^H$
  $\mathbf{a}_k^{[j]} = \sum_{n=1}^{N} \sigma_{d_{n,k}}^{-2[j]} \mathbf{X}_{n,k} x_{n,k}^{(1)*}$
  $\mathbf{G}_k^{[j]} = \mathbf{A}_k^{-1[j]} \mathbf{a}_k^{[j]}$
  $r_{n,k}^{[j]} = \mathbf{G}_k^{[j]H} \mathbf{X}_{n,k}$
  $d_{n,k}^{[j]} = x_{n,k}^{(1)} - r_{n,k}^{[j]}$
  $\sigma_{d_{n,k}}^{2[j+1]} = \max\{|d_{n,k}^{[j]}|^2, \epsilon\}$
- $\mathbf{G}_k^{[j]}$ is the desired reverberation prediction weight vector after convergence.

speech coefficients for all frames at frequency bin k, as represented by the vector $\mathbf{d}_k$, is given by

$$p(\mathbf{d}_k) = \prod_{n=1}^{N} p(d_{n,k}) = \prod_{n=1}^{N} \frac{1}{\pi \sigma_{d_{n,k}}^2} \exp\left(-\frac{|d_{n,k}|^2}{\sigma_{d_{n,k}}^2}\right) \tag{5}$$

Now, by inserting $d_{n,\,k}$ from (3) into (5), the joint distribution $p(\mathbf{d}_k)$ can be viewed as a function of the regression vector $\mathbf{G}_k$ and the desired speech spectral variances $\sigma_{\mathbf{d}_k}^2 = \{\sigma_{d_{1,k}}^2, \sigma_{d_{2,k}}^2, \cdots, \sigma_{d_{N,k}}^2\}$. Denoting this set of unknown parameters at each frequency bin by $\Theta_k = \{\mathbf{G}_k, \sigma_{\mathbf{d}_k}^2\}$ and taking the negative of logarithm of $p(\mathbf{d}_k) \equiv p(\mathbf{d}_k|\Theta_k)$ in (5), the objective function for Θ_k can be written as

$$\begin{aligned} \mathcal{J}(\Theta_k) &= -\log\ p(\mathbf{d}_k|\Theta_k) \\ &= \sum_{n=1}^{N} \left(\log\ \sigma_{d_{n,k}}^2 + \frac{\left|x_{n,k}^{(1)} - \mathbf{G}_k^H \mathbf{X}_{n,k}\right|^2}{\sigma_{d_{n,k}}^2}\right) \end{aligned} \tag{6}$$

where the constant terms have been discarded. To obtain the ML estimate of the parameter set Θ_k, (6) has to be minimized w.r.t. Θ_k. Since the optimization of (6) jointly w.r.t. $\mathbf{G}_k$ and $\sigma_{\mathbf{d}_k}^2$ is not mathematically tractable, an alternative sub-optimal solution is suggested in Nakatani et al. (2008b, 2010), where a two-step optimization procedure is performed, where at each step, only one of the two parameter subsets $\mathbf{G}_k$ and $\sigma_{\mathbf{d}_k}^2$ is optimized alternatively. The two-step procedure is repeated iteratively until a convergence criterion is satisfied or a maximum number of iterations is reached. While this strategy is rather straightforward, there is no guarantee that the alternating procedure results in a globally optimal solution (Jukic et al., 2015). A summary of the conventional WPE method is outlined in Table 1. It should be noted that, due to the simple instantaneous estimator used for $\sigma_{d_{n,k}}^2$, as seen in this table, the obtained value for this parameter has to be lower bounded by ϵ to avoid unreasonably small values when $|d_{n,\,k}|$ approaches zero.

In the following section, we propose an extension of the WPE approach by taking into account the correlation of $d_{n,\,k}$ across the STFT frame index, n, namely the IFC.

3. WPE method using inter-frame correlations

To demonstrate the importance of the temporal correlation in the desired early speech component, $d_{n,\,k}$, across STFT frames, which is the main motivation to develop the WPE method using IFC in this work, we have illustrated in Fig. 1 the IFC present in the early speech for a given frame lag. To generate this figure, we extracted the early part, i.e. the first 60 ms, of a room impulse response (RIR) with 60 dB reverberation time $T_{60\text{dB}} = 800$ ms, and then convolved it with the anechoic speech utterance to obtain the early speech $d_{n,\,k}$.[1] Next, the IFC measure $|E\{d_{n,k} d_{n-l,k}^*\}|$ was estimated through time averaging (i.e. long-term recursive smoothing) of the product $d_{n,k} d_{n-l,k}^*$ over n and then normalized by the estimated value of $E\{|d_{n,\,k}|^2\}$. The plotted values are the average over all frequency bins and have been obtained for the lag of $l = 3$. As observed from Fig. 1, the amount of correlation between the early speech components $d_{n,\,k}$ and $d_{n-l,k}$ is quite considerable as compared to the spectral variance $E\{|d_{n,\,k}|^2\}$. Whereas this correlation is neglected in earlier versions of the WPE method, the method that we here propose takes this correlation into account by jointly modeling the early speech terms. From Fig. 1, it is also observed that, even though the updating rate of the underlying smoothing is not high, the estimated IFC fluctuates rapidly across frames. Therefore, an efficient approach with fast convergence should be devised for its estimation.

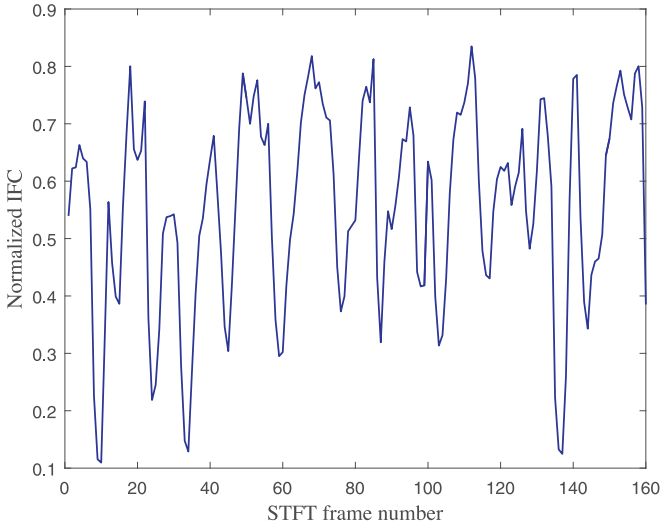

Fig. 1. Normalized IFC of the early speech $d_{n,\,k}$ averaged over frequency bins versus STFT frame number for a selected speech utterance.

In this section, we first derive a solution for the reverberation prediction vector $\mathbf{G}_k$ by considering the IFC, in contrast to the model in (5). Next, based on an extension of the method proposed for the estimation of the speech spectral variance in Parchami et al. (2016), an approach for the estimation of the IFC matrix of the desired speech terms, as required by the derived solution, will be developed.

3.1. Proposed approach

Considering the joint distribution of the desired speech STFT coefficients and assuming the independence across frequency bins, the temporally/spectrally independent model in (5) should be replaced by

$$p(\mathbf{d}_k) = p(d_{1,k}) \prod_{n=2}^{N} p(d_{n,k}|\mathbf{D}_{n,k}) \tag{7}$$

with $p(d_{n,\,k}|\mathbf{D}_{n,\,k})$ denoting the distribution of $d_{n,\,k}$ conditioned on $\mathbf{D}_{n,\,k} = [d_{n-1,k}, d_{n-2,k}, \cdots, d_{1,k}]^T$. Considering the fact that $d_{n,\,k}$ depends only on a limited number of the speech coefficients from previous frames, or equivalently, the fact that the IFC length is finite, (7) can be written as

[1] Note that, considering $D = 3$ early terms and using a frame length of 40 ms with 50% overlap, the early speech component corresponds to the first 60 ms of the RIR.

$$p(\mathbf{d}_k) = p(d_{1,k}) \prod_{n=2}^{N} p(d_{n,k}|\mathbf{d}'_{n-1,k}) = p(d_{1,k}) \prod_{n=2}^{N} \frac{p(d_{n,k}, \mathbf{d}'_{n-1,k})}{p(\mathbf{d}'_{n-1,k})} \tag{8}$$

where the conditioning term $\mathbf{D}_{n,k}$ in (7) has been replaced by the shorter segment $\mathbf{d}'_{n-1,k} = [d_{n-1,k}, d_{n-2,k}, \cdots, d_{n-\tau_k,k}]^T$ with τ_k as the assumed IFC length in frames. Unfortunately, proceeding with the model in (8) to find an ML solution for the regression vector $\mathbf{G}_k$ does not lead to a convex optimization problem. Therefore, to overcome this limitation, we alternatively exploit an approximate model by considering only the correlations among the frames within each segment, $\mathbf{d}'_{n,k} = [d_{n,k}, d_{n-1,k}, \cdots, d_{n-\tau_k+1,k}]^T$, and disregarding the correlations across the segments. This results in the following approximate model

$$p(\mathbf{d}_k) \simeq \prod_{n=1}^{\lfloor \frac{N}{\tau_k} \rfloor} p\left(\mathbf{d}'_{n,k}\right) = \prod_{n=1}^{\lfloor \frac{N}{\tau_k} \rfloor} \frac{1}{\pi^{\tau_k} \det \mathbf{\Phi}_{n,k}} \exp\left(-\mathbf{d}'^H_{n,k} \mathbf{\Phi}^{-1}_{n,k} \mathbf{d}'_{n,k}\right) \tag{9}$$

where $\mathbf{\Phi}_{n,k} = E\{\mathbf{d}'_{n,k}\mathbf{d}'^H_{n,k}\}$ represents the correlation matrix of $\mathbf{d}'_{n,k}$, det denotes the determinant of a matrix and $\lfloor . \rfloor$ is the floor function. Now, using (3), the desired speech segment $\mathbf{d}'_{n,k}$ can be expressed as

$$\mathbf{d}'_{n,k} = \mathbf{u}_{n,k} - \mathbf{U}^H_{n,k}\mathbf{h}_k \tag{10}$$

where

$$\mathbf{u}_{n,k} = [x^{(1)}_{n,k}, x^{(1)}_{n-1,k}, \cdots, x^{(1)}_{n-\tau_k+1,k}]^T \tag{11}$$
$$\mathbf{U}_{n,k} = [\mathbf{X}_{n,k}, \mathbf{X}_{n-1,k}, \cdots, \mathbf{X}_{n-\tau_k+1,k}]^*$$
$$\mathbf{h}_k = \mathbf{G}^*_k$$

In the same manner as the original WPE method (Nakatani et al., 2010), by considering the negative of the logarithm of $p(\mathbf{d}_k|\mathbf{h}_k)$, an ML-based objective function for the regression weight vector $\mathbf{h}_k$ can be derived as follows,

$$\mathcal{J}(\mathbf{h}_k) \triangleq -\log\ p(\mathbf{d}_k|\mathbf{h}_k) = \sum_{n=1}^{\lfloor \frac{N}{\tau_k} \rfloor} \left(\mathbf{d}'^H_{n,k} \mathbf{\Phi}^{-1}_{n,k} \mathbf{d}'_{n,k} + \mathcal{K}_{n,k}\right) \tag{12}$$

with $\mathcal{K}_{n,k}$ representing the terms independent of $\mathbf{h}_k$, which can be discarded. Inserting (10) into (12) and doing further manipulation result in

$$\mathcal{J}(\mathbf{h}_k) = \sum_{n=1}^{\lfloor \frac{N}{\tau_k} \rfloor} \left(\mathbf{h}^H_k \mathbf{A}_{n,k} \mathbf{h}_k - \mathbf{b}^H_{n,k}\mathbf{h}_k - \mathbf{h}^H_k \mathbf{b}_{n,k} + c_{n,k}\right) \tag{13}$$

where we defined

$$\mathbf{A}_{n,k} = \mathbf{U}_{n,k}\ \mathbf{\Phi}^{-1}_{n,k}\ \mathbf{U}^H_{n,k}$$
$$\mathbf{b}_{n,k} = \mathbf{U}_{n,k}\ \mathbf{\Phi}^{-1}_{n,k}\ \mathbf{u}_{n,k} \tag{14}$$
$$c_{n,k} = \mathbf{u}^H_{n,k}\ \mathbf{\Phi}^{-1}_{n,k}\ \mathbf{u}_{n,k}$$

Now by neglecting the constant term $c_{n,k}$, (13) can be arranged as

$$\mathcal{J}(\mathbf{h}_k) = \mathbf{h}^H_k \widetilde{\mathbf{A}}_k \mathbf{h}_k - \widetilde{\mathbf{b}}^H_k \mathbf{h}_k - \mathbf{h}^H_k \widetilde{\mathbf{b}}_k \tag{15}$$

with $\widetilde{\mathbf{A}}_k$ and $\widetilde{\mathbf{b}}_k$ as

$$\widetilde{\mathbf{A}}_k = \sum_{n=1}^{\lfloor \frac{N}{\tau_k} \rfloor} \mathbf{A}_{n,k}\ , \qquad \widetilde{\mathbf{b}}_k = \sum_{n=1}^{\lfloor \frac{N}{\tau_k} \rfloor} \mathbf{b}_{n,k} \tag{16}$$

It can be shown that the matrix $\widetilde{\mathbf{A}}_k$ is positive semidefinite, and therefore, the quadratic objective function in (15) is real-valued and convex in terms of $\mathbf{h}_k$. Subsequently, to find the global minimum of $\mathcal{J}(\mathbf{h}_k)$, we can express (15) in the following form

$$\mathcal{J}(\mathbf{h}_k) = \left(\mathbf{h}_k - \widehat{\mathbf{h}}_k\right)^H \widetilde{\mathbf{A}}_k \left(\mathbf{h}_k - \widehat{\mathbf{h}}_k\right) + c'_k \tag{17}$$

where c'_k is an independent term and

$$\widehat{\mathbf{h}}_k = \widetilde{\mathbf{A}}^{-1}_k \widetilde{\mathbf{b}}^H_k \tag{18}$$

It is evident that $\widehat{\mathbf{h}}_k$ in the above is the global minimum of the objective function $\mathcal{J}(\mathbf{h}_k)$ in (17), or equivalently, it is the estimate of the reverberation prediction weights by the proposed WPE method.

3.2. Estimation of the IFC matrix

To calculate the optimal reverberation prediction weights by (18), $\widetilde{\mathbf{A}}_k$ and $\widetilde{\mathbf{b}}_k$ in (16), and in turn, $\mathbf{A}_{n,k}$ and $\mathbf{b}_{n,k}$ given by (14) have to be calculated. To do so, as seen in (14), the IFC matrix of the desired speech terms, $\mathbf{\Phi}_{n,k}$, has to be estimated beforehand. In Parchami et al. (2016), a new variant of the WPE method has been suggested, that exploits the geometric spectral subtraction approach in Lu and Loizou (2008) along with the estimation of late reverberation spectral variance (LRSV), in order to estimate the spectral variance of the desired speech, $\sigma^2_{d_{n,k}}$, unlike the iterative scheme in the original WPE method, as in Table 1. We here develop an extension of the proposed method in Parchami et al. (2016) to estimate the spectral cross-variances of the desired speech terms, $\rho_{n_1,n_2,k} = E\{d_{n_1,k}d^*_{n_2,k}\}$, which in fact constitute the IFC matrix $\mathbf{\Phi}_{n,k}$. In this regard, by resorting to the dereverberation by spectral enhancement (gain function-based approach), the following estimate of $d_{n,k}$ can be obtained (Parchami et al., 2016)

$$\hat{d}_{n,k} = \sqrt{\frac{1 - \frac{(\gamma_{n,k} - \xi_{n,k} + 1)^2}{4\gamma_{n,k}}}{1 - \frac{(\gamma_{n,k} - \xi_{n,k} - 1)^2}{4\xi_{n,k}}}}\ x^{(1)}_{n,k} \tag{19}$$

where the two parameters $\xi_{n,k}$ and $\gamma_{n,k}$ are defined as

$$\xi_{n,k} = \frac{|d_{n,k}|^2}{|r_{n,k}|^2}, \quad \gamma_{n,k} = \frac{|x^{(1)}_{n,k}|^2}{|r_{n,k}|^2} \tag{20}$$

with $r_{n,k} = x^{(1)}_{n,k} - d_{n,k}$ as the reverberant-only component. We exploit (19) to provide primary estimates of $d_{n_1,k}$ and $d_{n_2,k}$ and then use recursive smoothing of $d_{n_1,k}d^*_{n_2,k}$ to estimate the elements of the IFC matrix $\mathbf{\Phi}_{n,k}$. As explained in Parchami et al. (2016), due to the unavailability of $|d_{n,k}|^2$ and $|r_{n,k}|^2$, the two parameters defined in (20) are not known *a priori* and have to be substituted by their approximations. To this end, we use $|\hat{d}_{n-1,k}|^2$ given by (19) for $|d_{n,k}|^2$ and a short-term estimate of the spectral variance $\sigma^2_{r_{n,k}}$ for $|r_{n,k}|^2$. To determine the spectral variance $\sigma^2_{r_{n,k}}$, we resort to the statistical model-based estimation of the LRSV, which has been widely explored in the context of spectral enhancement (Habets, 2007). Therein, an estimator of the LRSV was derived, using a statistical model for the RIR in the spectral domain along with a few recursive smoothing schemes. In brief, the following scheme has been conventionally used to estimate the LRSV (Habets et al., 2009)

$$\sigma^2_{x^{(1)}_{n,k}} = (1-\beta)\ \sigma^2_{x^{(1)}_{n-1,k}} + \beta|x^{(1)}_{n,k}|^2 \tag{21a}$$

$$\sigma^2_{\tilde{r}_{n,k}} = (1-\kappa)\ \sigma^2_{\tilde{r}_{n-1,k}} + \kappa\ \sigma^2_{x^{(1)}_{n-1,k}} \tag{21b}$$

$$\sigma^2_{r_{n,k}} = e^{-2\alpha_k R N_e}\ \sigma^2_{\tilde{r}_{n-(N_e-1),k}} \tag{21c}$$

where α_k is related to the 60 dB reverberation time, $T_{60\text{dB},\,k}$, by $\alpha_k = 3\log 10/(T_{60\text{dB},\,k} f_s)$ with f_s as the sampling frequency in Hz, R is the STFT frame advance in samples, β and κ are two smoothing parameters and N_e is the frame delay specifying the number of assumed early speech frames, which is chosen herein as D. This choice is made so that the number of frames considered as early speech in the LRSV estimation scheme will be equal to the number of included frames in the desired speech $d_{n,\,k}$ by the WPE method, as in (2). The term $\tilde{r}_{n,k}$ actually represents the entire reverberant speech including both the early and late reverberations but excluding the direct-path. Using the LRSV estimator in (21), the short-term estimate of $\sigma^2_{r_{n,k}}$ is obtained by choosing the smoothing parameters β and κ to be close to one. By this choice, the estimate of $\sigma^2_{r_{n,k}}$ includes more updates (less smoothing) and will therefore be a closer approximation to $|r_{n,\,k}|^2$. Yet, to avoid unreasonably small values for the approximated $|r_{n,\,k}|^2$ in the denominator of (20), this parameter is lower thresholded by 10^{-3}.

Now given the estimate of the desired speech components $\hat{d}_{n_1,k}$ and $\hat{d}_{n_2,k}$, as by (19), it is straightforward to use a recursive smoothing scheme to estimate the spectral cross-variance $\rho_{n_1,n_2,k}$, as the following

$$\widehat{\rho}_{n_1,n_2,k} = (1-\eta)\widehat{\rho}_{(n_1-1),(n_2-1),k} + \eta\ \hat{d}_{n_1,k}\hat{d}^*_{n_2,k} \tag{22}$$

with η as a fixed smoothing parameter. Equivalently, by expressing (22) in matrix form, it follows

$$\widehat{\mathbf{\Phi}}_{n,k} = (1-\eta)\widehat{\mathbf{\Phi}}_{n-1,k} + \eta\ \hat{\mathbf{d}}'_{n,k}\hat{\mathbf{d}}'^{H}_{n,k} \tag{23}$$

with the vector of the estimated desired speech terms $\hat{\mathbf{d}}'_{n,k} = [\hat{d}_{n,k}, \hat{d}_{n-1,k}, \cdots, \hat{d}_{n-\tau_k+1,k}]^T$. The inverse of the estimated IFC matrix $\widehat{\mathbf{\Phi}}_{n,k}$ is to be used to obtain $\mathbf{A}_{n,\,k}$ and $\mathbf{b}_{n,\,k}$ in (14). Here, to avoid the complexity involved in direct inversion of $\widehat{\mathbf{\Phi}}_{n,k}$ and also to overcome the common singularity issue encountered in the inversion of the sample correlation matrix, we use the Sherman–Morrison matrix inversion lemma (Hager, 1989) to implicitly invert $\widehat{\mathbf{\Phi}}_{n,k}$, as given by (23). The simplified form of this lemma can be written as (Hager, 1989)

$$\left(\mathcal{A} - \mathcal{U}\mathcal{V}^H\right)^{-1} = \mathcal{A}^{-1} + \frac{\mathcal{A}^{-1}\mathcal{U}\mathcal{V}^H\mathcal{A}^{-1}}{1-\mathcal{V}^H\mathcal{A}^{-1}\mathcal{U}} \tag{24}$$

for an invertible matrix $\mathcal{A}$ and any two column vectors $\mathcal{U}$ and $\mathcal{V}$. Using (24) for the inverse of $\widehat{\mathbf{\Phi}}_{n,k}$ in (23), i.e. by taking $\mathcal{A}$, $\mathcal{U}$ and $\mathcal{V}$ respectively as $(1-\eta)\widehat{\mathbf{\Phi}}_{n-1,k}$, $-\eta\hat{\mathbf{d}}_{n,k}$ and $\hat{\mathbf{d}}'_{n,k}$, it can be deduced that

$$\widehat{\mathbf{\Phi}}^{-1}_{n,k} = \frac{\widehat{\mathbf{\Phi}}^{-1}_{n-1,k}}{1-\eta} - \frac{\eta}{1-\eta}\ \frac{\widehat{\mathbf{\Phi}}^{-1}_{n-1,k}\ \hat{\mathbf{d}}'_{n,k}\ \hat{\mathbf{d}}'^{H}_{n,k}\ \widehat{\mathbf{\Phi}}^{-1}_{n-1,k}}{1-\eta+\eta\ \hat{\mathbf{d}}'^{H}_{n,k}\ \widehat{\mathbf{\Phi}}^{-1}_{n-1,k}\ \hat{\mathbf{d}}'_{n,k}} \tag{25}$$

The above can be recursively implemented to update the inverse of $\widehat{\mathbf{\Phi}}_{n,k}$ at each frame without the need for direct matrix inversion.

It should be noted that the overall WPE-based dereverberation approach presented in this section can be considered as an extension of the method presented in Parchami et al. (2016), by taking into account the IFC of the desired speech signal. Namely, for the choice of τ_k=1, it can be shown that the proposed solution in (18) degenerates to the method suggested in Parchami et al. (2016).

4. Performance evaluation

In this section, the performance of the proposed dereverberation approach is evaluated in comparison with the original WPE method and a few recent variations of this method from the literature. To this end, 20 clean speech utterances (including 10 male and 10 female speakers) are used from the TIMIT database (Garofolo et al., 1993), where the average length of the speech samples is 3.7 s and the average speech-to-silence ratio is 4.8. Both

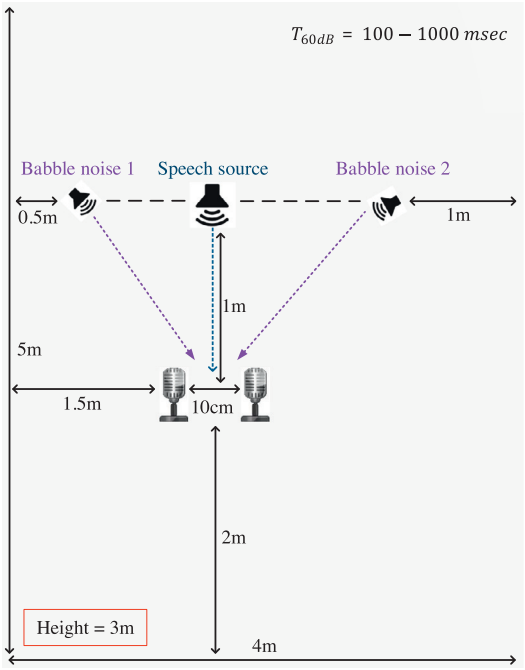

Fig. 2. A two-dimensional illustration for the geometry of the synthesized scenario of a noisy reverberant environment.

real-world recorded and synthetic RIRs are used to generate microphone array signals in a reverberant noisy environments. In the first case, to account for a real-world scenario, the clean speech utterances are convolved with measured RIRs from the SimData of the REVERB Challenge (REVERB, 2013), wherein an 8-channel circular microphone array with a diameter of 20 cm was placed in three rectangular rooms (labeled 1–3) to measure the RIRs.[2] Room 1 is 3.7 × 5.5 m with $T_{60\text{dB}}$ of 250 ms, room 2 is 4.8 × 6.2 m with $T_{60\text{dB}}$ of 680 ms and room 3 is 6.6 × 6.1 m with $T_{60\text{dB}}$ of 730 ms. The height for all rooms is 2.5 m and the microphone array and speakers are placed 1.1 m high. To account for different types of noise (i.e. babble, white and pink), the resulting reverberant signals are combined with different noises taken from the Noisex-92 (Varga and Steeneken, 1993) database at a global signal-to-noise ratio (SNR) of 15 dB. Although we report the results for three types of noise here, considering other types of noise led to the same conclusions as the ones drawn next. To properly add noise to the reverberant signals, we use the function *v_addnoise* of the speech processing toolbox VoiceBOX (Brookes, 2009), which calculates the speech signal level according to the ITU-T recommendation P.56 (ITU-T, 1993). In the second case, to further analyze the performance of the considered methods under different levels of reverberation, the image source method (ISM) (Lehmann, 2016) is used to simulate the scenario illustrated in Fig. 2. As viewed, a source of anechoic speech and two independent anechoic sources of babble noise taken from Noisex-92 (Varga and Steeneken, 1993) are placed in an acoustic room with the indicated dimensions. The RIRs from the speech and noise sources to the microphones are synthesized to achieve a desired reverberation time $T_{60\text{dB}}$. These are convolved with the anechoic signals to generate reverberant microphone signals, which are next linearly combined to achieve a desired global SNR of 15 dB.

For the relative evaluation of different dereverberation methods, we use four performance measures, as recommended by the REVERB Challenge (Kinoshita et al., 2013). These performance metrics

[2] Note that only two of the available 8 channels are used herein.

include: the perceptual evaluation of speech quality (PESQ), the cepstrum distance (CD), the frequency-weighted segmental SNR (FW-SNR) and the signal to reverberation modulation energy ratio (SRMR). The PESQ score is one of the most frequently used performance measures in the speech enhancement literature, which has been recommended by ITU-T standards for speech quality assessment (PESQ, 2001). It ranges between 1 and 4.5 with higher values corresponding to better speech quality. The CD is calculated as the log-spectral distance between the linear prediction coefficients (LPC) of the enhanced and clean speech spectra (Hu and Loizou, 2008). It is often limited in the range of [0,10], where a smaller CD value shows less deviation from the clean speech. The FW-SNR is calculated based on a critical band analysis with mel-frequency filter bank and using clean speech amplitude as the corresponding weights (Hu and Loizou, 2008). It generally takes a value in the range of [−10,35] dB with the higher the better. The SRMR, which has been exclusively devised for the assessment of dereverberation, is a non-intrusive measure (i.e., one requiring only the enhanced speech for its calculation), and is based on an auditory-inspired filterbank analysis of critical band temporal envelopes of the speech signal (Falk et al., 2010). A higher SRMR refers to a higher energy of the anechoic speech relative to that of the reverberant-only speech.

In the conducted experiments, the sampling rate is set to 16 kHz and a 40 ms Hamming window with overlap of 50% is used for the STFT analysis-synthesis. To achieve the best dereverberation performance, the number of early speech terms is chosen as $D = 3$ while the order of the regression vector $\mathbf{G}_k$ is chosen as $L_k = 20$. To implement the original (iterative) version of the WPE method (Nakatani et al., 2010), we take the number of iteration to be $J = 5$, since more iterations do not result in better performance. The number of microphones is taken to be $M = 2$, as the results obtained by using larger number of microphones lead to the same conclusions. We use the first 10 s of the reverberant speech observation to estimate the reverberation prediction weights $\mathbf{G}_k$ in all reported experiments. We take the length of IFC to be independent of frequency, i.e. $\tau_k \equiv \tau$, for our experiments.

In order to perform the matrix inversion in (18) with better accuracy, we use the QR factorization of the matrix $\widetilde{\mathbf{A}}_k$ in (16) with forward-backward substitution (Press et al., 2007). Also, to estimate the LRSV through (21), knowledge of the reverberation time $T_{60\text{dB}}$ is required. Here, we use the reverberation time estimation method in Löllmann et al. (2010) to estimate this parameter blindly from the observed speech. The estimated $T_{60\text{dB}}$ in this way is accurate enough not to degrade the performance of the underlying LRSV estimator in (21). The smoothing parameters β and κ in (21) are respectively selected as 0.5 and 0.8 while η in (25) is fixed at 0.7. Our approach requires no *prior* knowledge of the direct to reverberant ratio (DRR) parameter or its estimate.

To investigate the IFC present between early speech terms with different frame lags, we calculated the normalized IFC by sample averaging over all frequency bins and frames. The results are shown in Fig. 3 for both anechoic and reverberant speech signals with different values of the reverberation time. As seen, the IFC is quite pronounced for smaller lag values (say 5 or less), but decreases to a lower level for larger lags. We will take into account this observation in choosing the appropriate IFC length, τ, in the sequel. A more detailed study of the IFC in the STFT domain can be found in Cohen (2005).

Next, we study the effect of the assumed number of correlated speech frames, τ, on the overall performance of the proposed dereverberation approach. It was found that the choice of this parameter is more dependent on the number of early speech frames, D, than on other involved parameters, e.g. L_k and $T_{60\text{dB}}$. This theoretically makes sense since the parameter D determines the duration of the early reflections, and therefore, the IFC is controlled by D

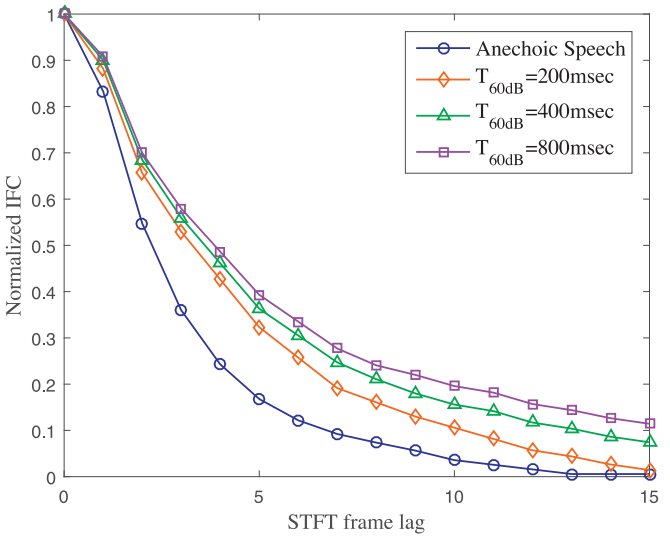

Fig. 3. Normalized IFC averaged over frequency bins and frames versus the frame lag for speech samples with different amounts of reverberation.

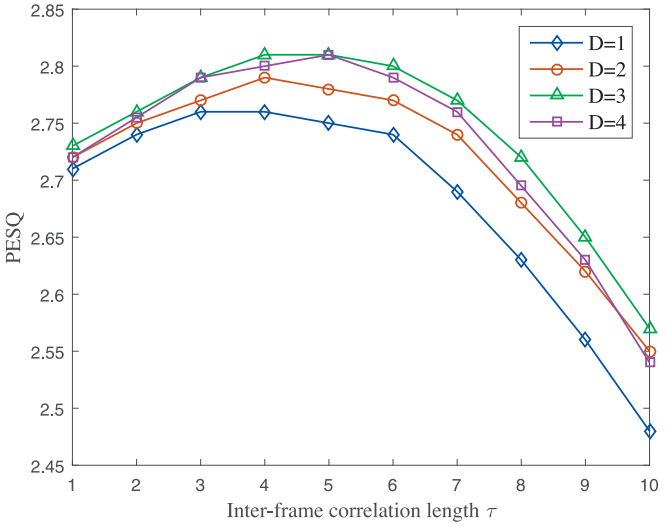

Fig. 4. Performance of the proposed WPE method versus the assumed IFC length, τ, for different D.

to a large extent. Fig. 4 shows the PESQ scores of the proposed approach versus different τ with D ranging from 1 to 4, when using the measured RIRs from the SimData of the REVERB Challenge. Apart from the observation that the performance of the proposed approach is best for $D = 3$, it can be seen that the higher the value of D the larger the value of the choice of τ resulting in the best performance. This result is due to the fact that the higher the value of D the larger the amount of the IFC between subsequent frames of the desired speech. It is also observed that the best choice of the parameter τ occurs in the range of 2-6, despite the fact that the theoretically optimal choice of τ is N, i.e., the number of frames in the entire speech utterance.[3] The reason for this limitation in the performance of the proposed approach seems to be due to the limited accuracy in the estimation of the IFC matrix, $\mathbf{\Phi}_{n,k}$. In effect, the estimation error in $\widehat{\mathbf{\Phi}}_{n,k}$, which grows with the size τ of the matrix $\mathbf{\Phi}_{n,k}$, degrades the overall performance of the proposed method. Therefore, we choose the value of $\tau = 5$ for the case of

[3] Note that in this case, the approximate model in (9) turns into an accurate joint model for all the desired speech frames.

Table 2
Performance comparison of different WPE-based dereverberation methods using the recorded RIR of room 1 from REVERB Challenge with babble noise.

Method	PESQ	CD	FW-SNR (dB)	SRMR (dB)
Unprocessed	2.26	4.26	2.90	3.82
Original WPE (Nakatani et al., 2010)	2.57	3.55	5.11	6.42
CGG-based WPE (Jukic et al., 2015)	2.60	3.50	5.33	6.74
WPE suggested in Parchami et al. (2016)	2.67	3.42	6.08	7.53
Proposed WPE	2.73	3.24	6.79	7.99
Proposed WPE with IFC knowledge	2.81	3.11	7.52	8.40

Table 3
Performance comparison of different WPE-based dereverberation methods using the recorded RIR of room 2 from REVERB Challenge with white noise.

Method	PESQ	CD	FW-SNR (dB)	SRMR (dB)
Unprocessed	1.94	4.62	0.90	2.05
Original WPE (Nakatani et al., 2010)	2.10	3.75	1.88	3.17
CGG-based WPE (Jukic et al., 2015)	2.12	3.70	1.98	3.25
WPE suggested in Parchami et al. (2016)	2.18	3.44	2.30	3.67
Proposed WPE	2.23	3.32	2.51	3.99
Proposed WPE with IFC knowledge	2.30	3.15	2.74	4.24

Table 4
Performance comparison of different WPE-based dereverberation methods using the recorded RIR of room 3 from REVERB Challenge with pink noise.

Method	PESQ	CD	FW-SNR (dB)	SRMR (dB)
Unprocessed	1.87	4.96	0.52	1.98
Original WPE (Nakatani et al., 2010)	2.01	3.82	1.38	3.09
CGG-based WPE (Jukic et al., 2015)	2.02	3.73	1.51	3.23
WPE suggested in Parchami et al. (2016)	2.07	3.50	1.82	3.60
Proposed WPE	2.13	3.36	2.06	3.87
Proposed WPE with IFC knowledge	2.21	3.19	2.29	4.20

$D = 3$ in our experiments. This is also consistent with the fact that the IFC is more strongly present in the lag values of around 5 or less, as inferred before from Fig. 3.

To evaluate the reverberation suppression performance of the proposed method, we compare it to the original WPE method (Nakatani et al., 2010), two recent developments of the same method based on the complex generalized Gaussian (CGG) family of distributions (Jukic et al., 2015) and the Laplacian distribution (Jukic and Doclo, 2014) for the desired speech, the WPE method using the estimation of speech spectral variance (Parchami et al., 2016), and finally, the proposed method by using the perfect knowledge of the desired speech component. The CGG-based method makes use of the same solution for the regression vector $\mathbf{G}_k$ as the original WPE method but with a different estimator of the speech spectral variance in its iterative procedure. The Laplacian-based method does not have a closed-form solution for the reverberation prediction weights, $\mathbf{G}_k$, and has to be implemented through numerical optimization, e.g. by using the CVX optimization toolbox (CVX Research, 2012). Next, the WPE method presented in Parchami et al. (2016), is actually a particular case of the presented method in this work by disregarding the IFC and estimating only the speech spectral variance at each frame independently. Finally, the proposed WPE method with IFC knowledge is obtained by exploiting only the early component of the speech (this can be obtained in the same manner as that for Fig. 1) as $\hat{\mathbf{d}}'_{n,k}$ in (23), and is considered as a reference for comparison. The comparative results obtained by using the recorded RIRs from REVERB Challenge with different noise types are presented in Tables 2–4 in terms of the aforementioned objective performance measures.

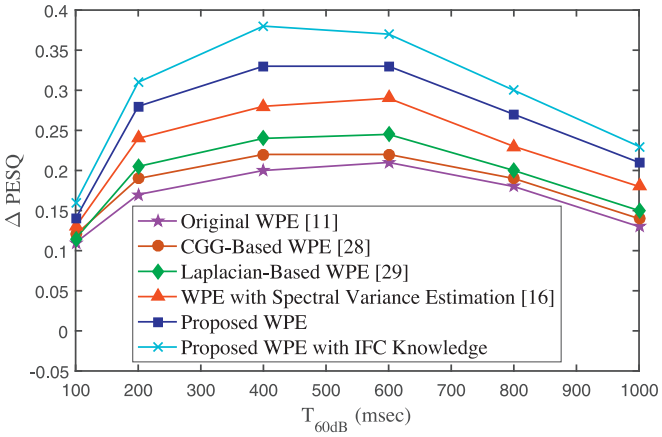

Fig. 5. Improvement in PESQ for different WPE-based dereverberation methods.

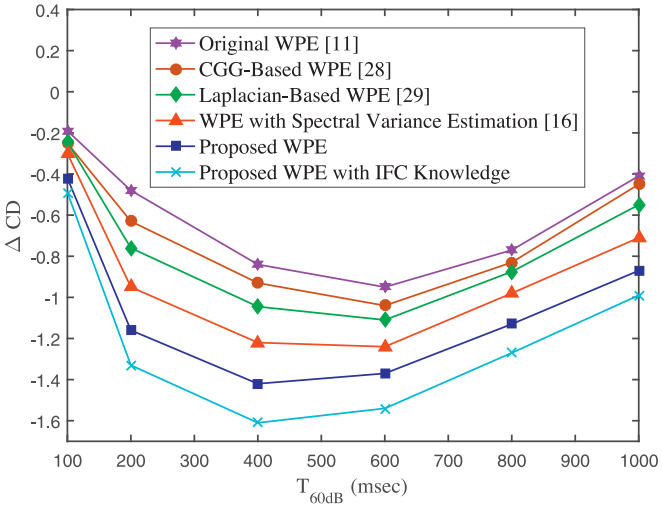

Fig. 6. Improvement in CD for different WPE-based dereverberation methods.

As observed, whereas the CGG and Laplacian-based methods achieve better scores w.r.t. the original WPE, the WPE with speech spectral variance estimation performs better than the former three methods, and finally, the proposed WPE method in this work achieves the best results as compared to the previous methods. It should be noted that the superior performance of the proposed WPE with knowledge of IFC shows the possibility of improving the proposed method through the availability of more accurate IFC matrix estimates. It is found that the relative performance of the considered methods in terms of the four investigated scores is consistent.

Next, to evaluate the performance of the considered dereverberation methods for different amounts of reverberation, the objective performance measures are obtained by using the synthesized RIRs with different $T_{60\text{dB}}$. The results are presented in Figs. 5–8 for $T_{60\text{dB}}$ in the range of 100–1000 ms. For better visualization, only the resulting improvements in the performance scores w.r.t. the unprocessed speech (denoted by ΔPESQ and such) are illustrated. As seen in these figures, the proposed method in this work and the one in Parchami et al. (2016), which are both based on the estimation of the speech spectral variance by means of an LRSV estimator, perform significantly better than the previous versions of the WPE method, which estimate the speech spectral variance iteratively along with the reverberation prediction weights. Also, it is observed that the proposed method achieves the best scores in comparison with the others in almost the entire range of $T_{60\text{dB}}$.

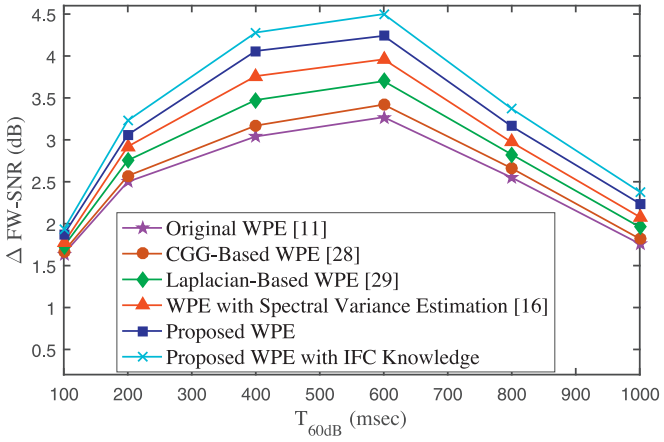

Fig. 7. Improvement in FW-SNR for different WPE-based dereverberation methods.

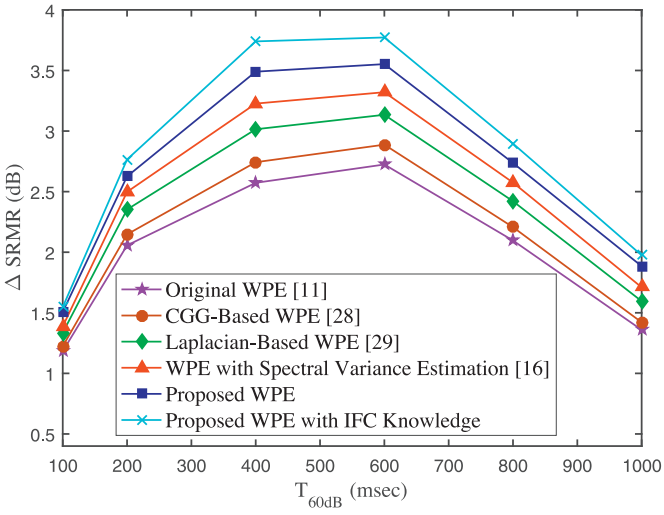

Fig. 8. Improvement in SRMR for different WPE-based dereverberation methods.

This advantage is more visible for the moderate values of T_{60dB}. There is also a considerable gap between the results obtained by using the proposed approach with the suggested estimation of the IFC matrix and those by using the perfect knowledge of the early speech, which indicates an avenue for further research for the estimation of the IFC.

5. Conclusion

In this work, we proposed a novel WPE dereverberation method based on an approximate model for the correlation across desired speech frames, namely the IFC, in the STFT domain. It was shown that, given an estimate of the IFC matrix, the dereverberation problem of interest can be formulated as a convex quadratic optimization leading to a closed-form Wiener-like solution. Performance evaluations using both recorded and synthesized RIRs reveal that the proposed method considerably outperforms the previous variations of the WPE method.

It can be concluded that incorporating the statistical model-based estimation of the desired speech spectral variance (or correlation matrix in general) into the WPE dereverberation method can lead to a better reverberation suppression performance. Such an approach, unlike the original WPE method, results in a non-iterative estimator for the reverberation prediction weight vector, provided that proper estimates of the spectral auto- and cross-variance of the desired speech terms are available. According to the performed experimentations, it can be concluded that the existing limit on the performance of the suggested WPE method in this work is mostly due to the limited accuracy in the estimation of the IFC matrix, and therefore, this shortcoming can be overcome by developing a more precise estimator for the IFC. This can serve as a topic of future research on linear prediction-based dereverberation in the STFT domain.

Acknowledgment

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Attias, H., Platt, J.C., Acero, A., Deng, L., 2001. Speech denoising and dereverberation using probabilistic models. Adv Neural Inf. Process. Syst. 13, 758–764.
Brookes, M., 2009. VoiceBOX: speech processing toolbox for MATLAB. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, last accessed on May 2016.
Cohen, I., 2005. Relaxed statistical model for speech enhancement and *a priori* SNR estimation. IEEE Trans. Speech Audio Process. 13 (5), 870–881.
CVX Research, I., 2012. CVX: matlab software for disciplined convex programming, version 2.0. Available at http://cvxr.com/cvx, last accessed on May 2016.
Erkelens, J., Heusdens, R., 2010. Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments. IEEE Trans. Audio Speech Language Process. 18 (7), 1746–1765.
Esch, T., 2012. Model-Based Speech Enhancement Exploiting Temporal and Spectral Dependencies. RWTH Aachen University Ph.D. thesis.
Falk, T.H., Zheng, C., Chan, W.Y., 2010. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Trans. Audio Speech Language Process. 18 (7), 1766–1774.
Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., Zue, V., 1993. TIMIT acoustic-phonetic continuous speech corpus LDC93S1. Philadelphia: Linguistic Data Consortium, last accessed on May 2016.
Habets, E.A.P., 2007. Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement. Technische Universiteit Eindhoven, Netherlands Ph.D. thesis.
Habets, E.A.P., Benesty, J., Chen, J., 2012. Multi-microphone noise reduction using interchannel and interframe correlations. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, pp. 305–308.
Habets, E.A.P., Gannot, S., Cohen, I., 2009. Late reverberant spectral variance estimation based on a statistical model. IEEE Signal Process. Lett. 16 (9), 770–773.
Hager, W.W., 1989. Updating the inverse of a matrix. SIAM Rev. 31 (2), 221–239.
Hu, Y., Loizou, P.C., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Language Process. 16 (1), 229–238.
Jukic, A., Doclo, S., 2014. Speech dereverberation using weighted prediction error with Laplacian model of the desired signal. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, pp. 5172–5176.
Jukic, A., van Waterschoot, T., Gerkmann, T., Doclo, S., 2015. Multi-channel linear prediction-based speech dereverberation with sparse priors. IEEE/ACM Trans. Audio Speech Language Process. 23 (9), 1509–1520.
Kinoshita, K., Delcroix, M., Nakatani, T., Miyoshi, M., 2009. Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. IEEE Trans. Audio Speech Language Process. 17 (4), 534–545.
Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., Maas, R., 2013. The reverb challenge: a common evaluation framework for dereverberation and recognition of reverberant speech. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, pp. 1–4.
Lehmann, E. A.,. Image-source method: matlab code implementationAvailable at http://www.eric-lehmann.com/, last accessed onMay 2016.
Löllmann, H.W., Yilmaz, E., Jeub, M., Vary, P., 2010. An improved algorithm for blind reverberation time estimation. In: Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC), Tel Aviv, Israel, pp. 1–4.
Lu, Y., Loizou, P.C., 2008. A geometric approach to spectral subtraction. Speech Commun. 50 (6), 453–466.
Nakatani, T., Juang, B.H., Yoshioka, T., Kinoshita, K., Delcroix, M., Miyoshi, M., 2008a. Speech dereverberation based on maximum-likelihood estimation with time–varying gaussian source model. IEEE Trans. Audio Speech Language Process. 16 (8), 1512–1527.
Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., Juang, B.H., 2008b. Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, USA, pp. 85–88.
Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., Juang, B.H., 2010. Speech dereverberation based on variance-normalized delayed linear prediction. IEEE Trans. Audio Speech Language Process. 18 (7), 1717–1731.

Naylor, P., Gaubitch, N. (Eds.), 2010, Speech Dereverberation. Springer-Verlag, London.
Parchami, M., Zhu, W.P., Champagne, B., 2016. Speech dereverberation using linear prediction with estimation of early speech spectral variance. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China.
Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 2007. Numerical recipes: the art of scientific computing (3rd ed.). New York: Cambridge University Press.
Recommendation P.56: Objective measurement of active speech level 1993. ITU-T.
Recommendation P.862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs 2001. ITU-T.
Schmid, D., Malik, S., Enzner, G., 2012. An expectation-maximization algorithm for multichannel adaptive speech dereverberation in the frequency-domain. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, pp. 17–20.
SimData: dev and eval sets based on WSJCAM0, 2013. REVERB challenge. Available at http://reverb2014.dereverberation.com/download.html, last accessed on May 2016.
Togami, M., Kawaguchi, Y., 2013. Noise robust speech dereverberation with Kalman smoother. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, pp. 7447–7451.
Varga, A., Steeneken, H.J.M., 1993. Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. 12 (3), 247–251.
Vaseghi, S.V., 2006. Advanced Digital Signal Processing and Noise Reduction. John Wiley & Sons. Chapter 17.
Yoshioka, T., Nakatani, T., Miyoshi, M., 2009. Integrated speech enhancement method using noise suppression and dereverberation. IEEE Trans. Audio Speech Language Process. 17 (2), 231–246.
Yoshioka, T., Sehr, A., Delcroix, M., Kinoshita, K., Maas, R., Nakatani, T., Kellermann, W., 2012. Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition. IEEE Signal Process. Mag. 29 (6), 114–126.