ELSEVIER
SPEECH
COMMUNICATION
An international journal of the European Association for Signal Processing (EURASIP) and of the International Speech Communication Association (ISCA)

Speech Reverberation Suppression for Time-Varying Environments Using Weighted Prediction Error Method With Time-Varying Autoregressive Model

Mahdi Parchami, Hamidreza Amindavar, Wei-Ping Zhu

PII: S0167-6393(18)30253-X
DOI: https://doi.org/10.1016/j.specom.2019.03.002
Reference: SPECOM 2630

To appear in: *Speech Communication*

Received date: 19 July 2018
Revised date: 4 March 2019
Accepted date: 10 March 2019

Please cite this article as: Mahdi Parchami, Hamidreza Amindavar, Wei-Ping Zhu, Speech Reverberation Suppression for Time-Varying Environments Using Weighted Prediction Error Method With Time-Varying Autoregressive Model, *Speech Communication* (2019), doi: https://doi.org/10.1016/j.specom.2019.03.002

SPEECH REVERBERATION SUPPRESSION FOR TIME-VARYING ENVIRONMENTS USING WEIGHTED PREDICTION ERROR METHOD WITH TIME-VARYING AUTOREGRESSIVE MODEL

Mahdi Parchami † *Hamidreza Amindavar* † *Wei-Ping Zhu* ⋆

† Department of Electrical Engineering
Amirkabir University of Technology, Tehran, Iran
{mparchami,hamidami}@aut.ac.ir

⋆ Department of Electrical and Computer Engineering
Concordia University, Montreal, Quebec, Canada
weiping@ece.concordia.ca

ABSTRACT

In this paper, a novel approach for the task of speech reverberation suppression in non-stationary (changing) acoustic environments is proposed. The suggested approach is based on the popular weighted prediction error (WPE) method, yet, instead of considering fixed reverberation prediction weights, our method takes into account the more generic time-varying autoregressive (TV-AR) model which allows dynamic estimation and updating for the prediction weights over time. We use an initial estimate of the prediction weights in order to optimally select the TV-AR model order and also to calculate the TV-AR coefficients. Next, by properly interpolating the calculated coefficients, we obtain the ultimate estimate of reverberation prediction weights. Performance evaluation of the proposed approach is shown not only for fixed acoustic rooms but also for environments where the source and/or sensors are moving. Our experiments reveal further reverberation suppression as well as higher quality in the enhanced speech samples in comparison with recent literature within the context of speech dereverberation.

***Index Terms*—** Dereverberation, speech enhancement, time-varying autoregressive model, weighted prediction error.

1. INTRODUCTION

Speech signals captured within an acoustic environment by distant microphones are subject to reflections from the surrounding surfaces, e.g., walls, ceiling and objects within the enclosure. This phenomenon, often referred to as reverberation, deteriorates the perceived quality/intelligibility of desired speech signals, and can also degrade to a large extent the performance of speech processing systems including hearing aids, hands-free teleconferencing, source separation and automatic speech recognition [1–3]. Therefore, efficient techniques for the suppression of reverberation in real-world acoustic environments is highly required in such applications.

Over the past two decades, there has been growing research on the development of various single- and multi-microphone (channel) reverberation suppression techniques. These techniques, known in the literature as speech dereverberation, can be broadly categorized into: blind system identification and inversion, multi-channel spatial processing, speech spectral enhancement and the probabilistic model-based approaches [2, 4]. Blind system identification methods aim at estimating the anechoic (clean) speech by processing reverberant observations by inverse filters that can be either calculated from available room impulse responses (RIRs) or estimated from the reverberant speech [5, 6]. More recent research in this direction, also termed as acoustic multi-channel equalization techniques, has been reported in [7, 8]. Within the class of multi-channel spatial processing, most conventional approaches exploit beamforming techniques to coherently combine the dominant early speech arrivals, as studied in [9, 10]. However, the dereverberation performance of beamforming methods is limited in general, unless a rather large number of microphones is used [2]. On the other hand, spectral enhancement (SE) methods, i.e., those based on applying a gain function on the corrupted speech[1], have also been employed for dereverberation [4]. The major advantage of the SE methods over the aforementioned techniques is their simplicity of implementation in the short-time Fourier transform (STFT) domain and the low computational complexity. More recent work in this direction can be found e.g. in [11, 12].

Another important category of dereverberation methods are the probabilistic model-based approaches leading to an optimal estimation of the anechoic speech in the statistical sense. In [13], probabilistic models of speech are incorporated into a variational Bayesian algorithm which estimates

[1] This category of methods was originally developed for the purpose of noise reduction, but later modified to perform reverberation suppression too.

the desired signal, the acoustic channels as well as the involved parameters in an iterative manner. A somewhat different strategy is followed in [14] where the parameters of all-pole models for both the desired speech and the reverberant components are iteratively determined by maximizing the likelihood function of the model parameters. Within this approach, a minimum mean-squared error (MMSE) estimator is derived that yields the enhanced speech. In a similar fashion, by using a time-varying statistical model for speech and a multi-channel linear prediction (MCLP) model for reverberation, efficient dereverberation methods have been derived in [15, 16]. Since the implementation of the aforementioned linear prediction-based methods in the time-domain is computationally costly, in [17, 18], a short-time Fourier transform (STFT) domain development of the MCLP-based dereverberation method is proposed. This method, referred to as the weighted prediction error (WPE) approach, is an iterative scheme that alternatively estimates the reverberation prediction coefficients and the anechoic speech spectral variance by using batch processing of the entire speech utterance.

Recently, a few variations of the WPE dereverberation method have been suggested and investigated in the relevant literature. In [19], instead of using the traditional Gaussian distribution for the desired speech, a Laplacian model is employed, which is known as a more accurate model. In [20], it is suggested to employ a general parametric sparse prior for the desired speech, which can be represented in a convex form as a maximization over scaled complex Gaussian distributions. Emphasizing the role of sparsity in speech dereverberation, the latter method is able to provide mild improvements over the conventional WPE method in most experiments. More recently, the authors in [21] consider modeling the temporal correlation across STFT coefficients, known as the inter-frame correlation (IFC), and exploit it in the derivation of the reverberation prediction weights. Thanks to the more realistic modeling used, the method is able to provide superior performance w.r.t. the previous literature within the tested scenarios. Finally, in [22], a constrained sparse version of the multi-channel WPE is proposed, wherein, in order to prevent overestimation of the undesired reverberant component, a statistical model is used for the estimation of late reverberation power spectral density (PSD). The consequent constrained optimization problem is solved therein by taking advantage of the alternating direction method of multipliers, resulting in a new variant of the WPE method.

In spite of its inherent advantages such as appealing performance, moderate complexity and not requiring prior information about the acoustic environment, the original WPE method and its variants still suffer from a few shortcomings. First, the corresponding reverberation prediction weights within this method are theoretically fixed w.r.t. STFT time frames. This lack of adaptation over time results in the prediction weights not being able to track changes happening in the acoustic environment while training the weights. In this sense, the same prediction weights are applied to the entire speech utterance and also the conventional method cannot be properly used in time-varying environments. Furthermore, this approach requires at least a few seconds of the observed speech utterance to ensure accurate convergence of the reverberation prediction coefficients, which may not be realistic in more online applications. Moreover, even though the WPE method is to some extent robust to background noise, basically, there is no solution suggested for handling the additive noise in the reverberant signal.

In this work, we employ a more accurate time-varying autoregressive (TV-AR) model for the reverberant speech, which takes into account the variability of the reverberation prediction weights within the training (batch) observations. Rather than considering the entire speech utterance, shorter segments of speech are used to estimate initial prediction weights using a modified version of the original WPE method. Next, the preliminary prediction weights are in turn exploited to form a TV-AR model for segments of speech. The ultimate prediction weights at each STFT frame are then estimated by properly interpolating across the TV-AR model coefficients over time frames.

This paper is outlined as follows. A brief review of the original WPE method along with the problem statement is presented in Section 2. Section 3 introduces the TV-AR model used withing the proposed reverberation prediction method. The proposed algorithm based on the WPE method is detailed in Section 4 while Section 5 is devoted to experimental results. Section 6 concludes this paper.

2. WPE METHOD: A BRIEF REVIEW

Suppose that a speech signal emitted from a single source is captured by M microphones placed in a reverberant enclosure. Considering the STFT-domain representation, let's denote the clean speech signal by $s_{n,k}$ with time frame index $n\in\{1,\ldots,N\}$ and frequency bin index $k\in\{1,\ldots,K\}$ where N is the total number of frames and K is the number of frequency bins. Using the linear prediction (LP) model, the reverberant speech signal observed at the m-th microphone, $x^{(m)}_{n,k}$, can then be represented in the STFT-domain [18]

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

where $h^{(m)}_{l,k}$ is an approximation of the acoustic transfer function (ATF) from the speech source to the m-th microphone, L_h denotes the length of the ATF (in frames) and $(.)^*$ denotes complex conjugation. The additive term, $e^{(m)}_{n,k}$, is the sum of LP error and the additive noise, and is neglected for simplicity as in [18]. Therefore, (1) can be written as

$$x^{(m)}_{n,k} = d^{(m)}_{n,k} + \sum_{l=D}^{L_h\text{-}1} h^{(m)*}_{l,k} s_{n-l,k} \tag{2}$$

where $d^{(m)}_{n,k} = \sum_{l=0}^{D-1} h^{(m)*}_{l,k} s_{n-l,k}$ is the sum of anechoic (direct-path) speech and early reflections at the m-th microphone and D is the duration of the early reflections. Most dereverberation techniques including the WPE method target the estimation of the desired signal, say $d_{n,k} \equiv d^{(1)}_{n,k}$, or equivalently, the suppression of the late reverberant terms represented by the summation in (2). Replacing the convolutive model in (2) by an autoregressive (AR) model gives the well-known MCLP form for the observation as follows

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

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

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

where $\mathbf{g}^{(m)}_{k}$ is the regression vector (reverberation prediction weights) of order L_k for the m-th channel and the superscript T is matrix transpose. The right-hand side of (3) is obtained by concatenating $\{\mathbf{x}^{(m)}_{n,k}\}$ and $\{\mathbf{g}^{(m)}_{k}\}$ over m to respectively form $\mathbf{X}_{n,k}$ and $\mathbf{G}_k$ both of length $L_k M$. Estimation of the regression vector, $\mathbf{G}_k$, and inserting it into (3) provides an estimate of the desired speech. From a statistical viewpoint, estimation of $\mathbf{G}_k$ can be performed by applying the maximum likelihood (ML) criterion at each frequency bin, k. In this sense, the conventional WPE method assumes a circularly symmetric complex Gaussian distribution for the desired speech coefficients, $d_{n,k}$, with (unknown) time-varying spectral variance, $\sigma^2_{d_{n,k}}{=}E\{|d_{n,k}|^2\}$, and zero mean [17,18]. Under the assumption that the desired speech STFT coefficients, $d_{n,k}$, are independent across frames, the joint distribution of the desired speech coefficients at frequency bin k, i.e., $\mathbf{d}_k$, is given by

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

Here, 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, $\sigma^2_{\mathbf{d}_k} = \{\sigma^2_{d_{1,k}}, \sigma^2_{d_{2,k}}, \cdots , \sigma^2_{d_{N,k}}\}$. Denoting the set of all unknown parameters $\Theta_k = \{\mathbf{G}_k, \sigma^2_{\mathbf{d}_k}\}$ and taking the negative logarithm of $p(\mathbf{d}_k) \equiv p(\mathbf{d}_k|\Theta_k)$ in (5), the objective (likelihood) function for Θ_k can be expressed as

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

where the constant terms have been discarded for ease of notation. To obtain the ML estimate of the parameter set, Θ_k, (6) has to be minimized w.r.t. Θ_k. Since the joint optimization of (6) w.r.t. $\mathbf{G}_k$ and $\sigma^2_{\mathbf{d}_k}$ is not mathematically tractable, an alternative suboptimal solution is followed in [18]. In this two-step procedure, (6) is optimized w.r.t. only one of the two parameter subsets, $\mathbf{G}_k$ or $\sigma^2_{\mathbf{d}_k}$, at each step and the two-step procedure is repeated iteratively until a convergence criterion is satisfied or a maximum number of iterations is reached. A summary of the explained conventional WPE method is outlined below as Algorithm 1.

Even though the original WPE method described above provides desirable performance in time-invariant reverberant environments, it has not been basically designed to cope with time-varying acoustic conditions, as will be investigated in Section 5. In the following section, we begin by introducing the time-varying prediction model, i.e. the TV-AR model, for the reverberant speech, and will next employ this model to derive the proposed reverberation prediction weights.

Algorithm 1: The conventional WPE method

- At each frequency bin k, consider the observations $x^{(m)}_{n,k}$, for all n and m, and the set of hyperparameters $\{D, L_k, \epsilon\}$.
- Initialize $\sigma^2_{d_{n,k}}$ by $\sigma^{2[1]}_{d_{n,k}}{=}|x^{(1)}_{n,k}|^2$.
- For, $j{=}1, 2, \cdots , J$ (with a fixed number of iterations, J) , repeat the following:

$$\begin{aligned} \mathbf{A}^{[j]}_k &= \textstyle\sum_{n=1}^{N} \sigma^{-2[j]}_{d_{n,k}} \mathbf{X}_{n,k} \mathbf{X}^H_{n,k} \\ \mathbf{a}^{[j]}_k &= \textstyle\sum_{n=1}^{N} \sigma^{-2[j]}_{d_{n,k}} \mathbf{X}_{n,k} x^{(1)*}_{n,k} \\ \mathbf{G}^{[j]}_k &= \mathbf{A}^{-1[j]}_k \mathbf{a}^{[j]}_k \\ r^{[j]}_{n,k} &= \mathbf{G}^{[j]H}_k \mathbf{X}_{n,k} \\ d^{[j]}_{n,k} &= x^{(1)}_{n,k} - r^{[j]}_{n,k} \\ \sigma^{2[j+1]}_{d_{n,k}} &= \max\{|d^{[j]}_{n,k}|^2, \epsilon\} \end{aligned}$$

- $\mathbf{G}^{[j]}_k$ is the desired reverberation prediction weight vector after J iterations.

3. TV-AR MODEL

Early studies on time-varying linear predictive models for speech began in [23] with the motivation that the human vocal tract often varies over time. It was proved that the so-called TV-AR model leads to increased accuracy in signal representation due to the continuously changing behavior of speech. More recent research in this direction was done e.g. in [24] with making use of a TV autoregressive moving average (ARMA) model for the purpose of covariance estimation. In [25], it was stated that such TV models are more efficient, since the inclusion of time variations in the model allows for analysis over longer data windows for speech processing. The TV-AR model can be used to derive a more flexible MCLP representation for reverberant speech as the following

$$d_{n,k} = x_{n,k}^{(1)} - \sum_{m=1}^{M} \mathbf{g}_{n,k}^{(m)H} \mathbf{x}_{n,k}^{(m)} \triangleq x_{n,k}^{(1)} - \mathbf{G}_{n,k}^{H} \mathbf{X}_{n,k} \quad (7)$$

where it is seen that the reverberation prediction weights, $\mathbf{G}_{n,k}$, are now a function of the STFT frame index, n, as opposed to (3). This is also in accordance with (2) where the ATF coefficients, $h_{l,k}$, are time-varying in the general sense. Considering the prediction weights, $\mathbf{G}_{n,k}$, to be changing with n, however, implies an infinite degree of freedom for the problem and thus makes the estimation of $\mathbf{G}_{n,k}$ tedious. In practice, the TV nature of the prediction weights is often modeled by choosing these coefficients as linear combinations of some known functions of time, namely, the basis functions [23,25]. With a model of this form, the LP weights, $\mathbf{G}_{n,k}$, can be expressed as a sum of Q coefficient vectors, $\mathbf{u}_q$, weighted by the basis functions, $f_q(n)$, as follows

$$\mathbf{G}_n = \sum_{q=0}^{Q-1} \mathbf{u}_q f_q(n) \quad (8)$$

It is seen that the frequency subscript k has been omitted for ease of notation. A few choices have been used in the literature for the set of known basis functions, $f_q(n)$, modeling the evolution of $\mathbf{G}_n$ with time. Popular candidates for speech applications include Legendre and Fourier polynomials, discrete prolate spheroidal functions, and even wavelets [25]. A suitable choice of this function set will be discussed later in Subsection 4.2. The prediction coefficients, $\mathbf{u}_q$, are in fact the coefficients of interest, which are to be estimated in a blind manner from the reverberant speech signal, as will be discussed in Section 4. In the rest of the current section, we explain our method of estimating the TV-AR model order, Q, as this is a matter of importance in the performance and accuracy of the proposed method.

3.1. Estimation of the Model Order

Order estimation for TV-AR models has been studied in a few works before. Existing methods include fixed empirical choices, ML and Bayesian estimation approaches [26]. In [27], an approach for model order estimation has been proposed for jointly Gaussian distributed data based on an accurate estimate of the observation covariance matrix. In here, due to the nature of the problem, we take a less restrictive ML-like approach as follows.

Suppose the entire reverberant speech utterance has been divided into segments of known length, R (in STFT frames), and the corresponding reverberation prediction weights have been initially estimated using the conventional WPE method at each segment[2]. Denoting the initial estimate of the prediction weights at segment λ by $\mathbf{G}_{\lambda}^{(0)}$, in a stochastic framework, the joint distribution of the initial prediction weights can be represented as

$$p(\mathbf{G}^{(0)}) = \prod_{\lambda=1}^{T} p(\mathbf{G}_{\lambda}^{(0)}) = \prod_{\lambda=1}^{T} p\left(\sum_{q=0}^{Q-1} \mathbf{u}_{\lambda,q} f_q(\lambda)\right) \quad (9)$$

where $T = \lfloor N/R \rfloor$ is the number of total segments with $\lfloor . \rfloor$ indicating the floor function, and we assumed independent estimates of the prediction weights across segments. Note that here, we considered the model of (8) at each speech segment, λ, and thus replaced n by λ, resulting in $\mathbf{G}_{\lambda}^{(0)}$ expressed as a weighted sum of coefficient vectors $\{\mathbf{u}_{\lambda,q}\}$ at each segment. Assuming zero-mean Gaussian distribution with independent identically distributed (i.i.d.) elements for the vector set $\{\mathbf{u}_{\lambda,q}\}$, the following can be deduced

$$\begin{aligned} p(\mathbf{G}_{\lambda}^{(0)}) = \prod_{\ell=1}^{LM} p(G_{\lambda_\ell}^{(0)}) = \prod_{\ell=1}^{LM} \frac{1}{\sqrt{2\pi\sigma_{\lambda}^2}} e^{-\frac{\left|G_{\lambda_\ell}^{(0)}\right|^2}{2\sigma_{\lambda}^2}} \\ = \frac{1}{(2\pi\sigma_{\lambda}^2)^{\frac{LM}{2}}} e^{-\frac{\sum_{\ell=1}^{LM}\left|G_{\lambda_\ell}^{(0)}\right|^2}{2\sigma_{\lambda}^2}} \end{aligned} \quad (10)$$

where $L{\equiv}L_k$, $G_{\lambda_\ell}^{(0)}$ denotes the ℓ-th element of $\mathbf{G}_{\lambda}^{(0)}$ and σ_{λ}^2 is the variance of the latter. Given that $G_{\lambda_\ell}^{(0)}{=}\sum_{q=0}^{Q-1} u_{\lambda,q,\ell} f_q(\lambda)$ from (9) and denoting the variance of each $u_{\lambda,q,\ell}$ term by $\sigma_{1,\lambda}^2$, we have for σ_{λ}^2

$$\sigma_{\lambda}^2 = \sum_{q=0}^{Q-1} \sigma_{1,\lambda}^2 \left|f_q(\lambda)\right|^2 = Q\, \sigma_{1,\lambda}^2 \quad (11)$$

where the right-hand side follows due to having basis functions with unit norm, e.g., complex Fourier coefficients. Next, inserting (10) into (9) by considering (11) gives

$$p(\mathbf{G}^{(0)}) = \frac{1}{\sqrt{(2\pi)^{LMT} \prod_{\lambda=1}^{T} \left(Q\sigma_{1,\lambda}^2\right)^{LM}}} e^{-\frac{1}{2}\sum_{\lambda=1}^{T} \frac{\left\|\mathbf{G}_{\lambda}^{(0)}\right\|_2^2}{Q\sigma_{1,\lambda}^2}} \quad (12)$$

where $||.||_2$ denotes the 2-norm of a vector. The logarithm of the expression in (12) given the unknown parameter Q can be viewed as the likelihood function of Q. Doing simple manipulations and discarding the constant terms, we get the following

$$\begin{aligned} \mathrm{LL}(Q, \sigma_{1,\lambda}^2) \triangleq -\log\left(p\left(\mathbf{G}^{(0)} \mid Q, \sigma_{1,\lambda}^2\right)\right) \\ \propto LM\left(T\log Q + \sum_{\lambda=1}^{T} \log \sigma_{1,\lambda}^2\right) + \frac{1}{Q}\sum_{\lambda=1}^{T} \frac{\left\|\mathbf{G}_{\lambda}^{(0)}\right\|_2^2}{\sigma_{1,\lambda}^2} \end{aligned} \quad (13)$$

with $\propto$ denoting equality but with omitting constant terms. Given an estimate of $\mathbf{G}_{\lambda}^{(0)}$, it is seen that the log-likelihood in the above is a function of both the model order Q and the

[2] Detailed explanation and reasoning of this strategy will be given in Section 4.

dummy variables $\{\sigma^2_{1,\lambda}\}$. Differentiating w.r.t. these variables and setting the result to zero gives the following set of equations

$$\widehat{\sigma}^2_{1,\lambda} = \frac{\left\| \mathbf{G}^{(0)}_{\lambda} \right\|^2_2}{LMQ}, \quad \lambda = 1, 2, \cdots, T$$
$$\widehat{Q} = \frac{1}{LMT} \sum_{\lambda=1}^{T} \frac{\left\| \mathbf{G}^{(0)}_{\lambda} \right\|^2_2}{\sigma^2_{1,\lambda}} \tag{14}$$

The two equations in (14) can be solved alternatively in an iterative manner by choosing an initial value for either of $\{\sigma^2_{1,\lambda}\}$ or Q. The rounded value of $\widehat{Q}$ to the closest integer at the end of iterations will be the ultimate estimate for the TV-AR model order.

4. PROPOSED ALGORITHM

In this section, we explain the proposed approach for the estimation of coefficients $\{\mathbf{u}_{\lambda,q}\}$, which leads to determination of the reverberation prediction weights $\mathbf{G}_{\lambda}$ through the TV-AR model as $\sum_{q=0}^{Q-1} \mathbf{u}_{\lambda,q} f_q(\lambda)$.

The functional expansion of (8) has been studied and applied to speech analysis previously in the literature and a few major methods have been adopted to track the coefficient trajectories, $\{\mathbf{u}_{\lambda,q}\}$, including the conventional least squares (LS) estimation followed in [23] or the stochastic filtering [28] and iterative methods [29]. In this work, due to the efficiency of the existing classic WPE method[3] in [18] and the restriction imposed by the number of available training observations within the online implementation, we instead make use of the initial estimate of the prediction weights used in the previous section, namely $\mathbf{G}^{(0)}_{\lambda}$, as well as a proper interpolation technique to determine the ultimate estimate for $\mathbf{G}_{\lambda}$.

Fig. 1 illustrates in order the steps of the proposed algorithm for the estimation of reverberation prediction weights, $\mathbf{G}_{\lambda}$. Note that the entire algorithm is applied independently to each frequency bin, k. As observed, first, the STFT frames of the reverberant speech observation, $\{\mathbf{x}^{(m)}_n\}$, are divided into segments of length R STFT frames. Next, we incorporate the total least squares (TLS) technique into the original WPE method in order to provide an initial estimate for the prediction weights, i.e. $\mathbf{G}^{(0)}_{\lambda}$, at each segment λ. The latter is exploited to estimate the TV-AR model order, i.e. $\widehat{Q}$, as discussed in Section 3.1. Having $\mathbf{G}^{(0)}_{\lambda}$ and $\widehat{Q}$ at hand, we next tend to estimate the $\widehat{Q}$ coefficient vectors of the TV-AR model, $\{\mathbf{u}_{\lambda,q}\}$. To this end, we first consider $\{\mathbf{u}_{\lambda,q}\}$ fixed over each set of $\widehat{Q}$ segments, namely, the block Λ, and use the $\widehat{Q}$ initially estimated prediction weights, $\mathbf{G}^{(0)}_{\lambda}$, in the TV-AR model to form a linear system of $LM\widehat{Q}$ equations/unknowns w.r.t. $\{\mathbf{u}_{\Lambda,q}\}$, as shown in Fig. 1. Solving this linear system and then using a proper interpolation technique to interpolate over the solution, $\widehat{\mathbf{u}}_{\Lambda,q}$, by a factor of $\widehat{Q}$ leads to the suggested estimate for $\mathbf{u}_{\lambda,q}$, i.e. $\widehat{\mathbf{u}}_{\lambda,q}$. The latter is in turn exploited in the TV-AR model, as seen in the figure, to come up with the ultimate estimate for the reverberation prediction weights at each segment, i.e. $\widehat{\mathbf{G}}_{\lambda}$. In the rest of this section, we explain in detail the aforementioned steps of the proposed algorithm in Fig. 1.

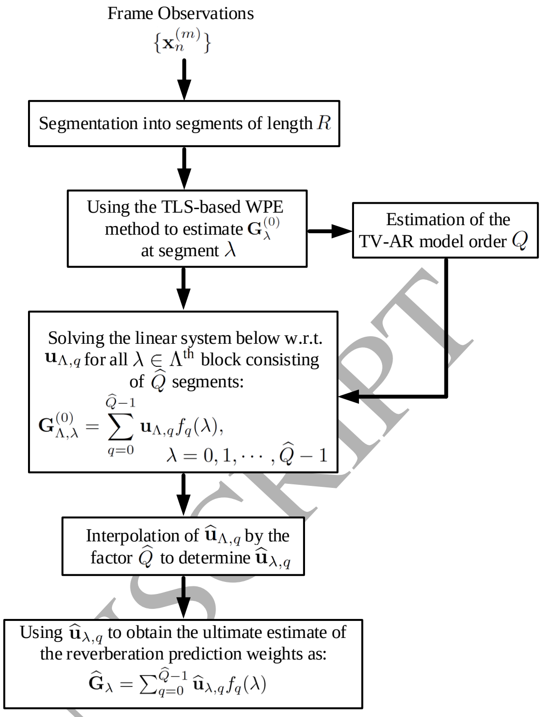
Frame Observations
$\{\mathbf{x}^{(m)}_n\}$
Segmentation into segments of length R
Using the TLS-based WPE method to estimate $\mathbf{G}^{(0)}_{\lambda}$ at segment λ
Estimation of the TV-AR model order Q
Solving the linear system below w.r.t. $\mathbf{u}_{\Lambda,q}$ for all $\lambda \in \Lambda^{\text{th}}$ block consisting of $\widehat{Q}$ segments:
$\mathbf{G}^{(0)}_{\Lambda,\lambda} = \sum_{q=0}^{\widehat{Q}-1} \mathbf{u}_{\Lambda,q} f_q(\lambda),$
$\lambda = 0, 1, \cdots, \widehat{Q}-1$
Interpolation of $\widehat{\mathbf{u}}_{\Lambda,q}$ by the factor $\widehat{Q}$ to determine $\widehat{\mathbf{u}}_{\lambda,q}$
Using $\widehat{\mathbf{u}}_{\lambda,q}$ to obtain the ultimate estimate of the reverberation prediction weights as:
$\widehat{\mathbf{G}}_{\lambda} = \sum_{q=0}^{\widehat{Q}-1} \widehat{\mathbf{u}}_{\lambda,q} f_q(\lambda)$

Fig. 1: Block diagram of the proposed algorithm for the estimation of reverberation prediction coefficients.

4.1. Estimation of the Initial Prediction Weights

The conventional WPE method [18] has been designed to estimate the reverberation prediction weights using the entire speech utterance as a batch. We here tend to employ this method for long enough segments of speech[4], indexed by λ, yet, we use a more efficient and generic approach to estimate the prediction weights at each segment of speech, namely, the TLS method.

Recall that in the WPE method, the cost function in (6) is minimized alternatively w.r.t. $\mathbf{G}$ and $\{\sigma^2_{d_n}\}$. In this alternation, suppose the minimization over $\mathbf{G}$ at each segment of

[3] This method is in fact based on the conventional LS technique in order to estimate the regression weights.

[4] A detailed analysis of this will be presented in Section 5.

length R in the following form

$$\min_{\mathbf{G}} \sum_{n=1}^{R} \left(\left| x_n^{(1)} - \mathbf{G}^H \mathbf{X}_n \right|^2 / \sigma_{d_n}^2 \right) = \min_{\mathbf{G}} \left\| \mathbf{x}^{\mathbf{1}} - \overline{\overline{\mathbf{X}}} \mathbf{G}^* \right\|_2^2 \qquad (15)$$

where we have

$$\begin{aligned} \mathbf{x}^{\mathbf{1}} &= \left[\frac{x_1^{(1)}}{\sigma_{d_1}}, \frac{x_2^{(1)}}{\sigma_{d_2}}, \ldots, \frac{x_R^{(1)}}{\sigma_{d_R}} \right]^T_{R\times 1} \\ \overline{\overline{\mathbf{X}}} &= \left[\frac{\mathbf{X}_1}{\sigma_{d_1}}, \frac{\mathbf{X}_2}{\sigma_{d_2}}, \ldots, \frac{\mathbf{X}_R}{\sigma_{d_R}} \right]^T_{R\times L} \end{aligned} \qquad (16)$$

The prediction weights, $\mathbf{G}_{LM\times 1}$, given by (15) are in fact the LS solution to the over-determined linear system $\overline{\overline{\mathbf{X}}}\mathbf{G}^* = \mathbf{x}^{\mathbf{1}}$ with $R > LM$. The latter can be obtained by using the pseudo-inverse of matrix $\overline{\overline{\mathbf{X}}}$, denoted by $\overline{\overline{\mathbf{X}}}^{\dagger}$, as

$$\widehat{\mathbf{G}}^* = \overline{\overline{\mathbf{X}}}^{\dagger} \mathbf{x}^{\mathbf{1}} = \left(\overline{\overline{\mathbf{X}}}^H \overline{\overline{\mathbf{X}}} \right)^{-1} \overline{\overline{\mathbf{X}}}^H \mathbf{x}^{\mathbf{1}} \qquad (17)$$

It can be easily shown that the estimate of $\mathbf{G}$ given by the original WPE method is equivalent to the one given by (17) at each speech segment. More precisely, due to the additive noise present in the observation $\mathbf{x}^{\mathbf{1}}$, the linear system can be considered to be of the form $\overline{\overline{\mathbf{X}}}\mathbf{G}^* = \mathbf{x}^{\mathbf{1}} + \widetilde{\mathbf{x}}^{\mathbf{1}}$ with $\widetilde{\mathbf{x}}^{\mathbf{1}}$ as an unknown perturbation term. In this regard, the LS solution in fact handles the aforementioned perturbation in the LS sense. However, the same uncertainty issue applies to the observation matrix $\overline{\overline{\mathbf{X}}}$ due to it consisting of noisy observation vectors. Hence, taking into account the more general problem given by $(\overline{\overline{\mathbf{X}}} + \widetilde{\mathbf{X}})\mathbf{G}^* = \mathbf{x}^{\mathbf{1}} + \widetilde{\mathbf{x}}^{\mathbf{1}}$ with $\widetilde{\mathbf{X}}$ as the perturbation of $\overline{\overline{\mathbf{X}}}$, we arrive at the TLS solution for $\mathbf{G}$. Next, according to theorem (2.6) in [30], following the basic solution to the TLS problem, we have

$$\widehat{\mathbf{G}}^* = \frac{-1}{V_{LM+1,LM+1}} [V_{1,LM+1}, V_{2,LM+1}, \cdots, V_{LM,LM+1}]^T \qquad (18)$$

where $V_{j,LM+1}$, $1 \leq j \leq LM+1$ is the j^{th} entry of the $(LM+1)^{\text{th}}$ column of the matrix $\mathbf{V}$ obtained from the singular value decomposition (SVD) of the matrix $[\overline{\overline{\mathbf{X}}}\ ,\ \mathbf{x}^{\mathbf{1}}]_{R\times(LM+1)}$, as following

$$[\overline{\overline{\mathbf{X}}}\ ,\ \mathbf{x}^{\mathbf{1}}] = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H \qquad (19)$$

The suggested solution for $\mathbf{G}$ in (18) is a more robust solution for the reverberation prediction, where the presence of noise in both observation arrays $\overline{\overline{\mathbf{X}}}$ and $\mathbf{x}^{\mathbf{1}}$ has been taken into account. Note that, similar to the original WPE method detailed as Algorithm 1, the speech variance terms, $\{\sigma_{d_i}\}_{i=1}^{R}$, in (16) are alternatively estimated along with the reverberation prediction weights, $\widehat{\mathbf{G}}^*$, in (18).

4.2. Determination of the Coefficient Vectors in the TV-AR Model

In this section, we present our approach to determining the TV-AR coefficient vectors, $\mathbf{u}_{\lambda,q}$, and thus the reverberation prediction weights $\mathbf{G}_{\lambda}$. In contrast with the state-of-the-art methods which include the least mean squared error (LMSE) estimators [23, 25], due to the limit on the number of available observations at each segment of speech, we take a different approach. In this sense, given each set, Λ, of $\widehat{Q}$ initial estimates of the prediction weights, $\mathbf{G}^{(0)}_{\Lambda,\lambda}$, we have the following

$$\mathbf{G}^{(0)}_{\Lambda,\lambda} = \sum_{q=0}^{\widehat{Q}-1} \mathbf{u}_{\Lambda,q} f_q(\lambda), \quad \lambda = 0, 1, \cdots, \widehat{Q}-1 \qquad (20)$$

which, in fact, is a linear system of $LM\widehat{Q}$ equations and unknowns w.r.t. $\{\mathbf{u}_{\Lambda,q}\}$. It should be noted that we here considered fixed coefficient vectors, $\mathbf{u}_{\Lambda,q}$, at each speech block, Λ, consisting of $\widehat{Q}$ segments. Arranging (20) in the matrix form results in the following

$$\overline{\overline{\mathbf{G}}}_{\Lambda} = \overline{\overline{\mathbf{U}}}_{\Lambda}\, \overline{\overline{\mathbf{F}}} \qquad (21)$$

where

$$\begin{aligned} \overline{\overline{\mathbf{G}}}_{\Lambda} &= [\mathbf{G}^{(0)}_{\Lambda,0}, \mathbf{G}^{(0)}_{\Lambda,1}, \cdots, \mathbf{G}^{(0)}_{\Lambda,\widehat{Q}-1}] \\ \overline{\overline{\mathbf{U}}}_{\Lambda} &= [\mathbf{u}_{\Lambda,0}, \mathbf{u}_{\Lambda,1}, \cdots, \mathbf{u}_{\Lambda,\widehat{Q}-1}] \\ \overline{\overline{\mathbf{F}}} &= [\mathbf{F}_0, \mathbf{F}_1, \cdots, \mathbf{F}_{\widehat{Q}-1}], \\ \mathbf{F}_{\lambda} &= [f_0(\lambda), f_1(\lambda), \cdots, f_{\widehat{Q}-1}(\lambda)]^T \end{aligned} \qquad (22)$$

Solving (21) w.r.t. $\overline{\overline{\mathbf{U}}}_{\Lambda}$ gives the following estimate for the TV-AR coefficient vectors

$$\widehat{\overline{\overline{\mathbf{U}}}}_{\Lambda} = \overline{\overline{\mathbf{G}}}_{\Lambda}\, \overline{\overline{\mathbf{F}}}^{-1} \qquad (23)$$

It is seen that the burden in calculating the solution to $\widehat{\overline{\overline{\mathbf{U}}}}_{\Lambda}$ is dominated by computing the inversion of matrix of basis functions, $\overline{\overline{\mathbf{F}}}$. Herein, by choosing the set of basis functions as discrete Fourier transform (DFT) bases, i.e., having

$$f_q(\lambda) = e^{-\frac{j2\pi\lambda q}{\widehat{Q}}}, \quad 0 \leq \lambda, q \leq \widehat{Q}-1 \qquad (24)$$

we can take advantage of the fast Fourier transform (FFT) algorithms in order to implement the matrix inversion, $\overline{\overline{\mathbf{F}}}^{-1}$, in (23).

Having at hand $\widehat{\overline{\overline{\mathbf{U}}}}_{\Lambda}$, or equivalently, the coefficient vectors $\widehat{\mathbf{u}}_{\Lambda,q}$ at each block, Λ, in the next step, we aim at interpolating $\widehat{\mathbf{u}}_{\Lambda,q}$ by the factor of $\widehat{Q}$ in order to obtain $\widehat{\mathbf{u}}_{\lambda,q}$, namely, the coefficient vectors at each segment λ. For this purpose, since there exists sparsity in the number of coefficients $\widehat{\mathbf{u}}_{\Lambda,q}$ compared to $\widehat{\mathbf{u}}_{\lambda,q}$, we choose to exploit some sparse polynomial interpolation technique. This setting allows us to use high order interpolating polynomials when having a smaller number of available interpolation nodes than the polynomial's order [31]. In essence, the polynomial interpolation problem of interest consists of fitting a sparse polynomial, i.e. one with many zero coefficients, to the given interpolation nodes and obtain the non-zero coefficients of the underlying polynomial. More elaborately, denoting the interpolating polynomial by $\mathcal{F}(\chi) = \sum_{j=0}^{\mathcal{D}} \omega_j \chi^j$, the problem of interest can be expressed as determining the coefficients $\{\omega_j\}_{j=0}^{\mathcal{D}}$ through fitting the polynomial to the interpolated values, $\widehat{\mathbf{u}}_{\Lambda,q}$, at the interpolation nodes, $\chi = 1 + \eta\widehat{Q}$ with $\eta \in \mathbb{Z}$. Doing so results in the following equation set

$$\boldsymbol{\Psi}_{q,\ell} = \boldsymbol{\Phi}\ \boldsymbol{\Omega}_{q,\ell} \tag{25}$$

where we have defined $\boldsymbol{\Psi}_{q,\ell} \triangleq [\widehat{u}_{1,q,\ell}, \widehat{u}_{2,q,\ell}, \cdots, \widehat{u}_{\mathcal{N},q,\ell}]^T$ with $\widehat{u}_{\Lambda,q,\ell}$ as the ℓ^{th} element of $\widehat{\mathbf{u}}_{\Lambda,q}$ for $1 \le \ell \le LM$ and $1 \le \Lambda \le \mathcal{N}$ with $\mathcal{N} = \left\lfloor T/\widehat{Q} \right\rfloor$ denoting the number of total speech blocks. Furthermore, $\boldsymbol{\Phi}_{\mathcal{N}\times(\mathcal{D}+1)}$ is the interpolation matrix with its Λ^{th} row being

$$\boldsymbol{\varphi}_\Lambda = [1, \alpha_\Lambda, \alpha_\Lambda^2, \cdots, \alpha_\Lambda^{\mathcal{D}}], \quad 1 \le \Lambda \le \mathcal{N} \tag{26}$$

with $\alpha_\Lambda = 1 + (\Lambda - 1)\widehat{Q}$ as the interpolation nodes, and $\boldsymbol{\Omega}_{q,\ell} = [\omega_{0,q,\ell},\ \omega_{1,q,\ell},\ \cdots,\ \omega_{\mathcal{D},q,\ell}]^T$ with $\omega_{j,q,\ell}$ denoting the j^{th} coefficient of the interpolating polynomial of interest for $0 \le q \le \widehat{Q} - 1$ and $1 \le \ell \le LM$.

In solving (25) for $\boldsymbol{\Omega}_{q,\ell}$, it should be noted that due to the limit on the number of blocks, $\mathcal{N}$, and the requirement for having a high enough polynomial order, $\mathcal{D}$, (25) often turns to an underdetermined set of equations with $\mathcal{N} < \mathcal{D} + 1$, and therefore, cannot be solved directly. With this in mind, we here make use of the compressive sensing theory [32] which presents a theoretical framework for investigating the sparse interpolation problem, assuming that the interpolating polynomial, $\mathcal{F}(\chi)$, is $\mathcal{S}$-sparse, meaning that there are only $\mathcal{S}$ non-zero entires in $\boldsymbol{\Omega}_{q,\ell}$ with $\mathcal{S} < \mathcal{D} + 1$. To this end, as discussed in the relevant literature [33, 34], one can use the ℓ_1-minimization approach to find the sparse solution to (25) with various settings for $\boldsymbol{\Phi}$, as the following

$$\widehat{\boldsymbol{\Omega}}_{q,\ell} = \underset{\boldsymbol{\Omega}_{q,\ell}}{\operatorname{argmin}} \ \ \|\boldsymbol{\Omega}_{q,\ell}\|_{\ell_1} \ \ \text{subject to } \boldsymbol{\Psi}_{q,\ell} = \boldsymbol{\Phi}\ \boldsymbol{\Omega}_{q,\ell} \tag{27}$$

where $\|.\|_{\ell_1}$ denotes the ℓ_1-norm. Solving the above gives the coefficients of the interpolating polynomial, $\mathcal{F}_{q,\ell}(\chi)$, which can be evaluated at $\chi \in \mathbb{N}$ to calculate the TV-AR coefficients, $\widehat{\mathbf{u}}_{\lambda,q}$. Based on this estimate of $\mathbf{u}_{\lambda,q}$, as our approach suggests in Fig. 1, we resort to the TV-AR model, $\widehat{\mathbf{G}}_\lambda = \sum_{q=0}^{Q-1} \widehat{\mathbf{u}}_{\lambda,q} f_q(\lambda)$, in order to obtain the ultimate estimate of the reverberation prediction weights at each speech segment, $\widehat{\mathbf{G}}_\lambda$. Parameter settings for the interpolation technique, e.g. the sparsity level $\mathcal{S}$, along with other parameter choices will be discussed in Section 5.

5. EXPERIMENTAL RESULTS

In this section, we investigate the dereverberation performance of the proposed approach in comparison with the original WPE method and a few recent variations of this method from the literature. Our evaluations are performed in both time-invariant and time-varying environments.

5.1. Implementation Details

For the evaluation of the reverberation suppression methods under study, we exploit anechoic (clean) speech utterances including 20 male and 20 female speakers from the TIMIT database [35] with the entire length of 20 sec. Here, the anechoic speech utterances are convolved with either the synthesized or measured RIRs, and next, noise samples are added to them. In our simulations, the sampling frequency, f_s, is set to 16 kHz and a 20 msec Hamming window with an overlap of 75% is used for the STFT analysis-synthesis. To implement our approach, as per Fig. 1, we consider dividing the entire speech utterance into segments of length R=40 STFT frames, and the estimation of the TV-AR model order, Q, discussed in Section 3.1 resulted in values typically in the range of [5,15] for the values scenarios under test. It should be noted that there exists a trade-off in choosing the length of segments, R, since too short segments may cause erroneous/unstable prediction weights, $\mathbf{G}_\lambda^{(0)}$, whereas a long segment length requires long speech utterances as the input and also slows down the rate of the adaptation of the prediction weights to a changing environment. To achieve the best performance with all the methods, the following parameter setting is used as per Algorithm 1: $\{D, L, \epsilon, J\}$=$\{3, 15, 10^{-3}, 5\}$. Further, as with the proposed approach in Section 4, the order of the interpolating polynomial, $\mathcal{D}$, and the sparsity level, $\mathcal{S}$, are respectively set to 15 and 5. The latter choice of the parameters revealed the best performance, considering the number of speech blocks, $\mathcal{N}$, which in turn depends on the length of the entire speech utterance under test. Further, unless otherwise stated, the number of microphones is taken as M=2. The results obtained by using a larger number of microphones led to similar conclusions. We use both synthetic and recorded (real-world) RIRs to generate reverberant noisy microphone array signals. The setup of the regarding scenarios will be explained in detail in subsections 5.2 and 5.3.

For the evaluation of the reverberation suppression of the methods under test, we use a few of the most frequent performance metrics recommended by REVERB Challenge [36], including 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 metrics in the speech enhancement literature and is the one recommended by ITU-T standards for speech quality assessment [37]. It often ranges between 1 and 4.5 with the higher values the better speech quality. The CD can be calculated as the log-spectral distance between the linear prediction coefficients (LPC) for the spectrum of the reverberant/enhanced and clean speech signals [38]. It is often limited in the range of [0,10], with a smaller value showing less deviation from the clean speech. The FW-SNR is calculated based on a critical band analysis with the mel-frequency filter bank with the clean speech amplitude as the corresponding weights [38]. The FW-SNR generally takes a value in the range of [-10,35] dB with the higher the better speech quality. The SRMR, which has been exclusively devised for the assessment of dereverberation, is a non-intrusive measure[5] and is based on an auditory-inspired filter bank analysis of the critical band temporal envelopes of speech [39]. A higher SRMR associates to a higher energy in the anechoic speech component relative to that of the reverberant-only component.

To evaluate the reverberation suppression performance of the proposed approach in Section 4, we compare it to the original WPE method [18], two more recent developments of this method based on the complex generalized Gaussian (CGG) family of distributions for the desired speech [20] and the Laplacian distribution for the desired speech [19], the adaptive sparse WPE method in [22], and finally, the WPE method using the inter-frame correlation (IFC), namely, the IFC-based WPE [21]. The CGG-based method basically makes use of the same solution as the original WPE method for the regression vector, $\mathbf{G}_k$, but with a different estimator of the speech spectral variance within the iterative procedure discussed in Section 2. The Laplacian-based method does not lead to a closed-form solution for $\mathbf{G}_k$ and has to be implemented through numerical optimization, e.g. by using the CVX optimization toolbox [40]. The latter is also used to handle the ℓ_1-minimization problem encountered in our approach, as in (27). The adaptive sparse WPE method uses a statistical model for the estimation of the reverberation spectral variance instead of alternatively estimating it like Algorithm 1 and tends to solve the problem by using the alternating direction method of multipliers. We employed the steps presented as Algorithm 3 in Section IV of [22]. Finally, the IFC-based WPE takes into account the inherent temporal correlations across STFT frames in developing a closed-form solution for $\mathbf{G}_k$, and in fact, extends the conventional solution to a more robust one.

At the end of this subsection, it should be noted that, for all the experiments discussed in the following subsections, having performed the reverberation suppression by different methods, we average the resulting performance measures over various speech files in order to deduce more reliable and consistent results.

5.2. Experiments with Synthetic RIRs

In order to analyze the performance of all methods under controllable levels of reverberation, the image source method (ISM) [41] is used to simulate different RIRs, as illustrated in Fig. 2. As seen, in this scenario with fixed geometry, a source of anechoic speech and a source of noise extracted from Noisex-92 database [42] have been placed in an acoustic room with the indicated dimensions. The RIRs from the speech and noise sources to the microphone array have been synthesized to achieve a 60 dB reverberation time in the range of 100 msec< T_{60dB} <1000 msec. The RIRs are then convolved with the corresponding anechoic speech files from TIMIT and noise files from Noisex-92 to generate reverberant microphone array signals. In this sense, a few different types of noise from Noisex-92 database were exploited to conduct the experiments, yet, results for the most challenging noise type, i.e., the babble noise, are reported here. We used the same noise file for different speech utterances or RIRs in the experiments. We consider a global reverberant signal-to-noise ratio (RSNR) of 15 dB for the scenario of Fig. 2, whereas different RSNR values for the scenarios with recorded RIRs. To properly add noise to the reverberant signals, we use the function *v_addnoise* from a speech processing toolbox, VOICEBOX [43], which calculates the speech signal level according to the ITU-T recommendation P.56 [44].

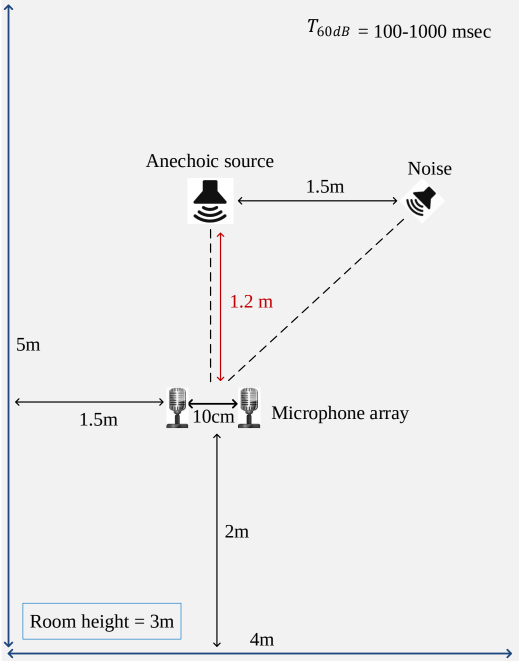

Fig. 2: Illustration of the devised experiment for the generation of time-invariant RIRs with the ISM method.

Furthermore, to investigate the performance of the considered dereverberation methods in time-varying environments, we set up the scenario in Fig. 3 with the ISM method to generate time-varying RIRs. As viewed, a talker is moving from the initial point at t=0 to the end point at t=20 seconds along the shown straight line, resulting in a time-varying RIR for the source-to-microphone channel. Herein, we approximate this continuous trajectory by 20 discrete points and determine the corresponding RIR at each point by using the ISM method. Next, a 20 sec anechoic speech utterance

[5] A non-intrusive measure is one requiring only the distorted/enhanced speech for its calculation.

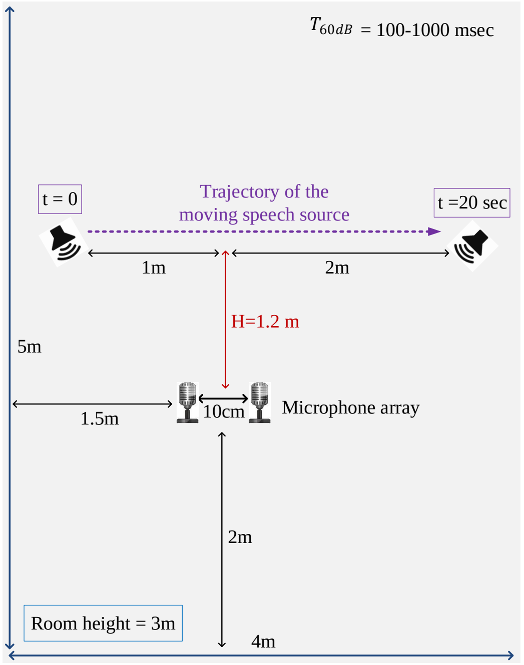

Fig. 3: Illustration of the devised experiment for the generation of time-varying RIRs with the ISM method.

is segmented into 20 segments and the resulting segments are filtered by the generated RIRs at the discrete points. The entire reverberant speech is then generated by concatenating the 20 individual segments into one. Please note that, in order to avoid unnatural changes happening in between the processed speech segments, we used a 50% overlap when segmenting and concatenating the entire speech file.

We first assess the performance of the suggested method for the estimation of the TV-AR model order, Q, explained in subsection 3.1. In this respect, the improvement in the aforementioned objective performance measures are obtained for the scenario of Fig. 2 when our approach is used with different values of Q as well as the estimated one. The corresponding results are shown in Fig. 4 for the PESQ metric versus T_{60dB}. For better visualization, only the resulting improvements w.r.t. the unprocessed speech (denoted by Δ PESQ) are illustrated. Whereas the estimated value for Q is 8 in this scenario, we also evaluated the performance for several other choices of Q. It can be seen that our log-likelihood method based on the estimation of the initial weights, $\mathbf{G}_{\lambda}^{(0)}$, due to its capability to adapt the value of $\hat{Q}$ to the reverberant speech signal, is able to provide significantly better performance, as compared to the other choices of Q. Whereas the value of $\widehat{Q}$ is consistent for the synthetic RIRs when using the ISM method with the same room geometry for both time-invariant and time-varying cases, different values of $\widehat{Q}$ are obtained when testing our approach under different room geometries or with different recorded (real-world) RIRs. The same result holds true when using other objective performance measures than PESQ.

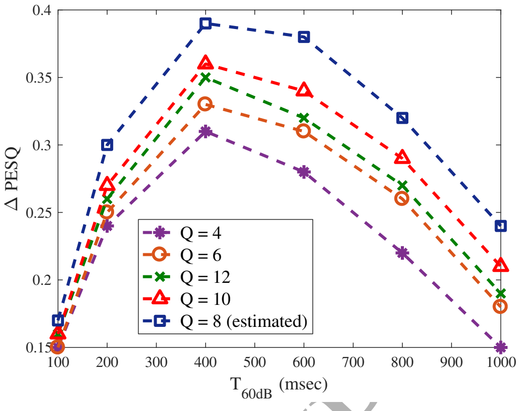

Fig. 4: Performance of the proposed approach with different values for the TV-AR model order, Q, in terms of the PESQ metric.

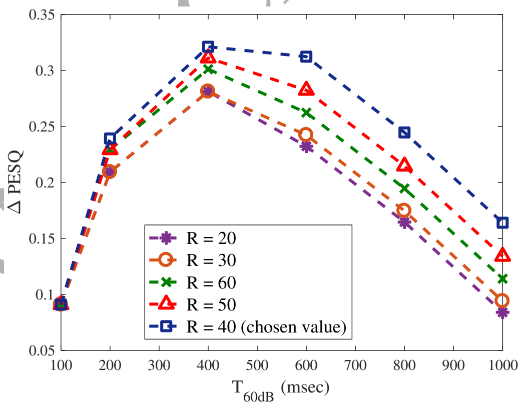

Fig. 5: Performance of the proposed approach with different values for the segment length, R, in terms of the PESQ metric.

Next, to determine the role of choosing the segment length, R, on the performance of the proposed method, we evaluate Δ PESQ scores using different choices of R for the time-varying scenario of Fig. 3. As seen in the results shown in Fig. 5, the chosen value of R=40 STFT frames results in the best possible performance. In fact, there exists a compromise in choosing the segment length, R. In this sense, a too short segment length results in unacceptably erroneous prediction weights, $\mathbf{G}_{\lambda}^{(0)}$, while a larger R reduces the rate of adaptation of the estimated $\mathbf{G}_{\lambda}^{(0)}$ to a changing RIR. It should be noted that the visible difference in the performance with changing R seen in Fig. 5 only appears in the experiments with time-varying RIRs and the aforementioned difference is negligible in time-invariant RIRs. Using other performance measures led to the same conclusion.

To investigate the performance of the considered dereverberation methods with T_{60dB}, we illustrate the four objective performance measures obtained by using different methods in Fig. 6 under the scenario of Fig. 2. As seen, the proposed

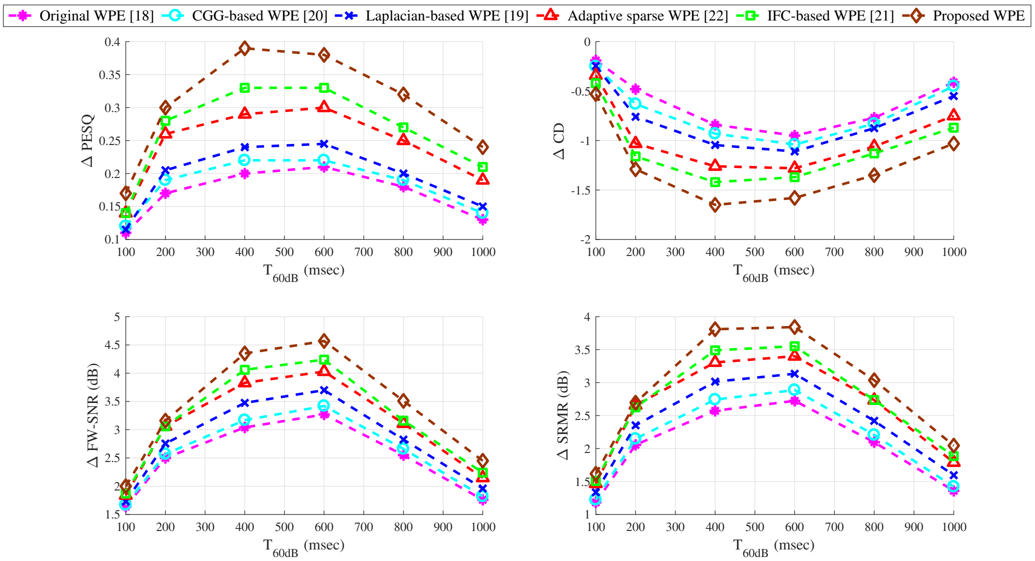

Fig. 6: Performance metrics obtained by using different WPE-based methods using the time-invariant scenario in Fig. 2.

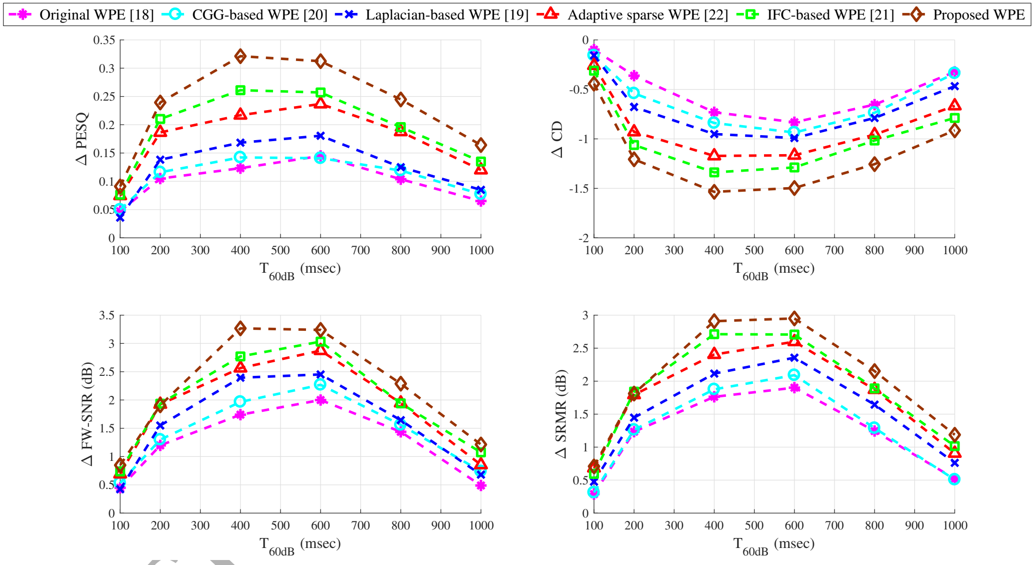

Fig. 7: Performance metrics obtained by using different WPE-based methods using the time-varying scenario in Fig. 3.

method in this work achieves considerably better scores than the previous versions of the WPE method over the entire range of T_{60dB}. The main reason for such performance advantage in the time-invariant case is the use of appropriate interpolation across the TV-AR coefficient vectors at each block, i.e. $\{\mathbf{u}_{\Lambda,q}\}$, to obtain the ultimate estimate of the reverberation prediction weights, $\mathbf{G}_{\lambda}$. Note that, within other WPE-based methods, only some averaging-like smoothing is performed to obtain the regression vector, $\mathbf{G}_{\lambda}$, and thus, increasing the length of the input training (batch) speech does not necessarily improve the precision of the prediction weights for long utterances. While we observed no considerable improvements with training utterances longer than 10 sec for the previous WPE-based methods, our approach was able to highly outperform the state-of-the-art methods for such reverberant speech samples. Furthermore, it is observed that this advantage is more visible for the moderate values of T_{60dB} ranging in the middle of the interval. This is due to the fact that, whereas the improvement in dereverberation is not pronounced for very small amounts of reverberation, for very heavily reverberant environments, the WPE method is able to provide slight improvements. Still, there exists considerable benefit with using the proposed approach in Section 4 as compared to the other methods. It was found in our experiments that the relative performance of the considered methods w.r.t. the four performance metrics from [36] is consistent.

We further perform the experiment shown in Fig. 3 to

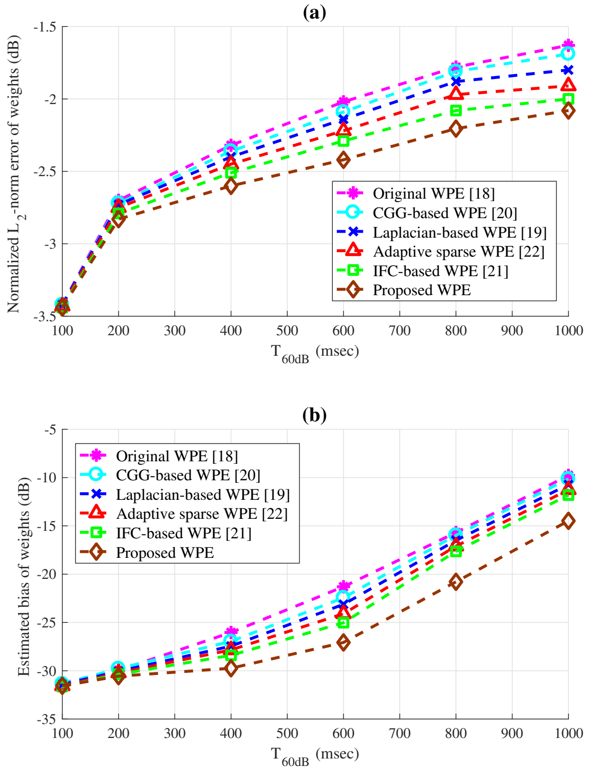

Fig. 8: Error analysis of the prediction weights for different methods w.r.t. the asymptotic weights, (a): Normalized error by (28), (b): Bias given by Jackknife resampling method.

assess the performance of the methods under study in a time-varying RIR. The corresponding results are illustrated in Fig. 7. As viewed, even though the average gain in the performance metrics is smaller in comparison with the time-invariant case, a trend similar to that in Fig. 6 holds true herein with the proposed approach being superior to the other methods in almost all T_{60dB}'s. Regarding the superiority of the suggested approach, the same reasoning as that in case of a time-invariant RIR, i.e., the use of proper interpolation along coefficient vectors, applies here. In addition, contrary to the state-of-the-art methods, our approach uses a TV-AR model for the reverberant speech, which causes the estimated prediction weights, $\mathbf{G}_\lambda$, to be adapted to a changing environment.

Next, to investigate how close the reverberation prediction weights, $\mathbf{G}$, are w.r.t. an asymptotically optimal set of prediction weights, namely $\mathbf{G}_{\text{opt}}$, we devise an experiment for the error analysis of $\mathbf{G}$ w.r.t. the latter. To this end, we calculate the normalized ℓ_2-norm of the error between $\mathbf{G}$ and $\mathbf{G}_{\text{opt}}$ as well as the existing bias between them for the case of time-invariant RIRs. The former error measure is defined as

$$\Delta(\mathbf{G}_{\text{opt}}, \mathbf{G}) = \underset{k}{E}\left\{\frac{||\mathbf{G}_{\text{opt}} - \mathbf{G}||_2^2}{||\mathbf{G}_{\text{opt}}||_2^2}\right\} \tag{28}$$

with $\mathbf{G}_{\text{opt}}$ as the asymptotically optimal prediction weights and $\underset{k}{E}\{.\}$ denoting the expectation (average) over the frequency bins, k. Further, we compute the statistical bias between $\mathbf{G}$ and $\mathbf{G}_{\text{opt}}$ by using the Jackknife resampling method [45] implemented by the 'Jackknife' function of Matlab, wherein we use the error $\mathbf{G} - \mathbf{G}_{\text{opt}}$ as the input vector and take the mean of the output magnitudes. Within this experiment, we choose a long reverberant speech utterance (of the length 30 sec) and use the original WPE method on the entire speech sample to obtain $\mathbf{G}_{\text{opt}}$. Yet, to compute the prediction weights, $\mathbf{G}$, for the methods under study, we consider only 4 sec segments from the entire utterance. By this way, since $\mathbf{G}_{\text{opt}}$ can act as an asymptotically optimal set of prediction weights for a time-invariant RIR, the aforementioned error measures between $\mathbf{G}$ and $\mathbf{G}_{\text{opt}}$ show how fast the prediction weights can adapt to the reverberant environment and the anechoic speech source. The corresponding two error measures are indicated in Fig. 8 versus T_{60dB}. As observed, the proposed approach clearly attains smaller error measures especially for high values of T_{60dB}. This advantage is even more visible in terms of the discussed measure of bias. It can be concluded that our interpolation-based method is able to provide a more robust behavior in online scenarios where only short speech utterances are available to process.

In order to investigate the advantage of the proposed approach in different conditions, we change the two important parameters, i.e., the source-to-microphone distance and the number of microphones, within the receiver array and measure the resulting performance metrics. The obtained results for the scenario of Fig. 2 are presented in Fig. 9 and Fig. 10, respectively. In these figures, the values indicated by `'ref'` refer to the reverberant (unprocessed) speech. In addition to the superiority of the proposed method clearly seen in these figures, it can be observed that there exists considerable performance improvement with increasing the number of microphones, especially from one to two.

Next, to measure the performance gain obtained by using the individual subsystems of the proposed algorithm, we devise two other experiments with different values of T_{60dB}. In the former, we employ the original WPE method to obtain the initial estimate of reverberation prediction weights, $\mathbf{G}_\lambda^{(0)}$, and compare the performance to that where the TLS-based WPE is used for initialization. The corresponding PESQ results have been averaged over T_{60dB} and shown in Tables 1 and 2, respectively for the experiments depicted in Fig. 2 and Fig. 3. Therein, Algorithm (A) refers to the proposed approach with initialization using the TLS-based WPE discussed in Section 4.1 and Algorithm (B) refers to that with initialization using original WPE. It is seen that, while both methods achieve very close scores in high RSNR values such as 20 dB, for lower RSNR values, using method (A) leads to considerable improvement w.r.t. method (B). This shows that the TLS-based method used for initialization is beneficial mostly in making the proposed approach robust w.r.t. the background noise in the input speech signal, preventing further performance degradation in presence of noise.

In the latter experiment, we evaluate the pure improve-

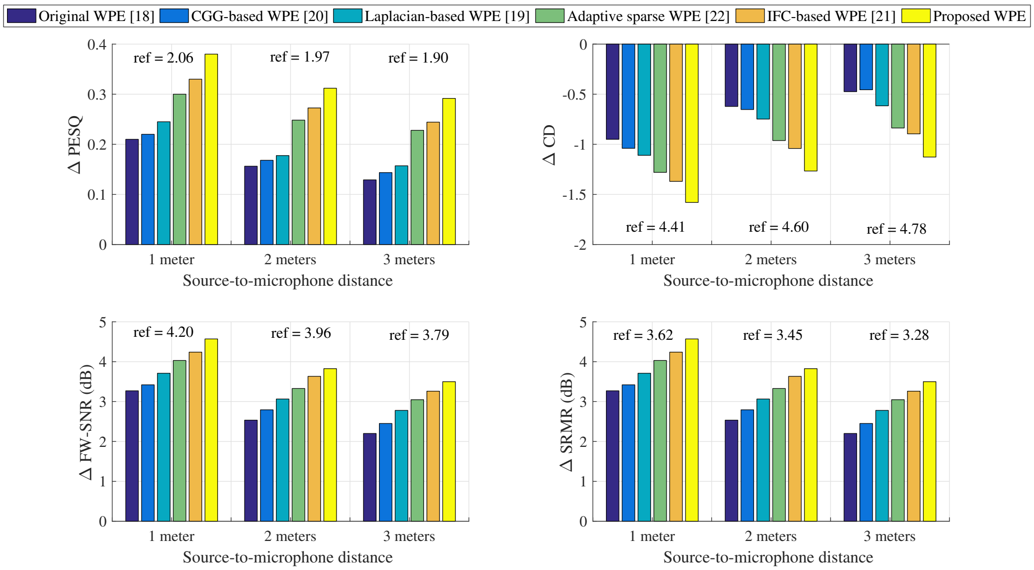

Fig. 9: Performance metrics obtained by using different WPE-based methods for several source-to-microphone distances.

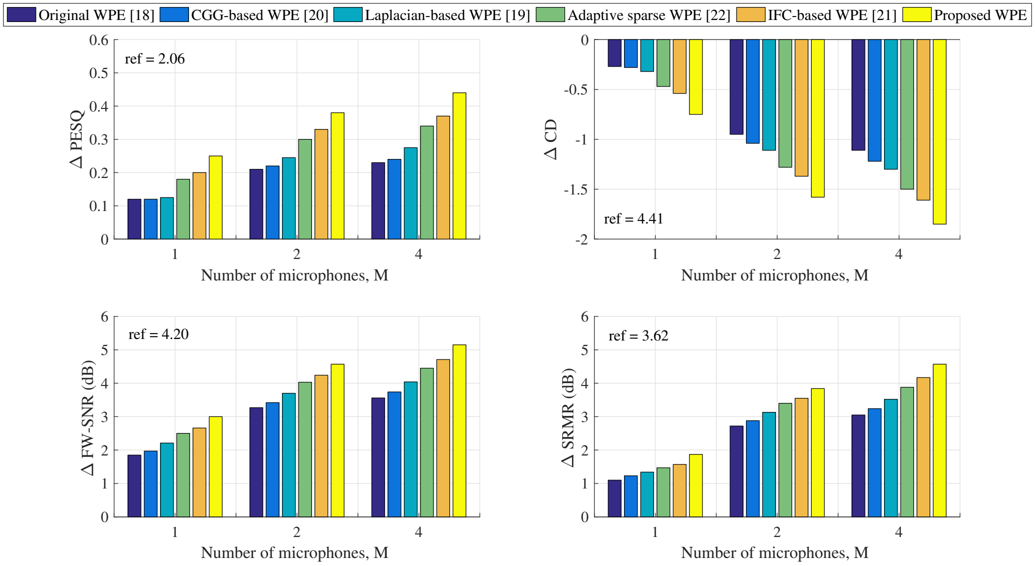

Fig. 10: Performance metrics obtained by using different WPE-based methods for different number of microphones, M.

Table 1: PESQ scores for Algorithm (A) and Algorithm (B), averaged over $T_{60dB} \in [100, 1000]$ msec in case of time-invariant RIRs.

RSNR	5 dB	10 dB	15 dB	20 dB
Unprocessed	2.02	2.12	2.33	2.49
Method (A)	2.16	2.37	2.70	2.93
Method (B)	2.09	2.33	2.67	2.93

Table 2: PESQ scores for Algorithm (A) and Algorithm (B), averaged over $T_{60dB} \in [100, 1000]$ msec in case of time-varying RIRs.

RSNR	5 dB	10 dB	15 dB	20 dB
Unprocessed	1.98	2.08	2.29	2.45
Method (A)	2.09	2.31	2.59	2.81
Method (B)	2.04	2.28	2.57	2.81

ment given only by making use of the TLS solution for reverberation prediction weights in (18), as compared to the LS solution employed in the original WPE discussed in Section 2. For doing so, we use the Algorithm 1 but with the prediction weights obtained by (18) and call this approach the TLS-based WPE. We then compare the results with the orig-

Table 3: PESQ scores for different WPE-based methods averaged over $T_{60dB} \in [100, 1000]$ msec in case of time-invariant RIRs.

RSNR	5 dB	10 dB	15 dB	20 dB
Unprocessed	2.02	2.12	2.33	2.49
Original WPE	2.08	2.26	2.52	2.73
TLS-based WPE	2.24	2.40	2.65	2.83
Proposed WPE	2.29	2.44	2.69	2.87

Table 4: PESQ scores for different WPE-based methods averaged over $T_{60dB} \in [100, 1000]$ msec in case of time-varying RIRs.

RSNR	5 dB	10 dB	15 dB	20 dB
Unprocessed	1.98	2.08	2.29	2.45
Original WPE	2.02	2.16	2.41	2.65
TLS-based WPE	2.10	2.23	2.47	2.69
Proposed WPE	2.18	2.32	2.58	2.80

inal WPE and the proposed WPE in Section 4, as presented in Tables 3 and 4. By careful inspection of the PESQ scores, it is inferred that there exists a consistent improvement obtained by exploiting the TLS technique over the LS technique employed by the original WPE method, especially in the lower RSNR values, indicating that the TLS approach is more advantageous in noisier conditions. Further, the increment in the resulting PESQ values obtained by the proposed approach in Section 4 w.r.t. those obtained by the TLS-based WPE, as observed in Tables 3 and 4, shows the pure performance advantage provided by the TV-AR model as well as the interpolation method discussed in Section 4.2, especially when dealing with time-varying environments.

5.3. Experiments with Recorded RIRs

To perform experiments in real-world time-invariant environments, the anechoic speech is convolved with measured RIRs from the SimData of the REVERB Challenge [46]. Therein, an 8-channel circular microphone array with a diameter of 20 cm was placed in three rectangular rooms (labeled as 1-3) to measure the RIRs[6]. Room 1 is 3.7 m×5.5 m with T_{60dB} of 250 msec, room 2 is 4.8 m×6.2 m with T_{60dB} of 680 msec and room 3 is 6.6 m×6.1 m with T_{60dB} of 730 msec. The height of all rooms is 2.5 m and the microphone array and speakers were placed 1.1 m high. In all scenarios with recorded RIRs in time-invariant environments, the reported results are the average among the three different rooms.

Furthermore, to demonstrate the advantage of the proposed approach in real world time-varying environments, we use the RevDyn database of recorded reverberant speech files [47]. Therein, the recordings were performed in a room with dimensions of 6 m×5.9 m×2.3 m and a T_{60dB} of 750 msec. There are 4 English speakers, namely, 2 females and 2 males, with each speaker performing 4 different experiments. The first 2 experiments involve speaking in different locations in the room and walking naturally between them. The next 2 experiments consist of only slight movements, e.g., head turning, sitting down and standing up. Each of the 4 experiments consist of three different scenarios, wherein each scenario is 1 min long (net speaking time is about 45 sec in each scenario). Therefore, the total number of recordings is $4 \times 12 = 48$ (1 min. each), with 9 channels (channel 9 is the reference microphone). The speaker-to-microphone distance varies between 2 m and 3.8 m and 8 omni-directional AKG-CK32 microphones are used to perform the recordings. Herein, we consider the 45 sec net speaking time of the recorded speech at each experiment and the presented values for the metrics are actually the average over the different scenarios. To take into account the effect of the background noise, we also add babble noise to the recorded reverberant signals at an RSNR in the range of [5,20] dB.

In Fig. 11, the improvement in terms of the four metrics are demonstrated for the case of time-invariant recorded RIRs. Babble noise utterances with the same features as those explained in Section 5.2 are added to the reverberant speech files. As viewed, in the middle to high RSNR values, our method is able to provide superior performance w.r.t. the less recent methods. The reason for decaying the improvement seen by all the WPE-based dereverberation methods for lower RSNRs is that, in essence, these methods have been designed to only cope with noiseless reverberant speech signals, and therefore, their performance degrades in adverse noisy conditions. Even though the same phenomenon happens to some extent with our approach, the benefit provided by the proposed approach is still relatively high in such RSNRs. The significant reason for this is the use of TLS solution for the prediction weights, $\mathbf{G}$, as in (18), which makes the solution robust w.r.t. the additive background noise by considering perturbations for the prediction weights. In a similar fashion to the above, Fig. 12 shows the averaged performance metrics obtained by using the dereverberation methods but for the time-varying recorded RIRs. It is observed that, in spite of smaller improvements compared to the time-invariant scenario of Fig. 11, the proposed approach clearly outperforms the less recent methods, particularly at higher RSNR values.

Next, using the same setting as that in Fig. 8, we analyze the two measures of error between the prediction weights by different methods and the asymptotically optimal one. In Fig. 13, the normalized ℓ_2-norm of the error and the statistical bias obtained by Jackknife method are shown for the case of time-invariant recorded RIRs. As viewed, especially at high RSNR values where the reverberation is dominant to the background noise, the proposed method results in smaller error measures and is therefore closer to the namely optimal prediction weights.

[6] Note that only two of the available 8 channels are used herein given $M = 2$.

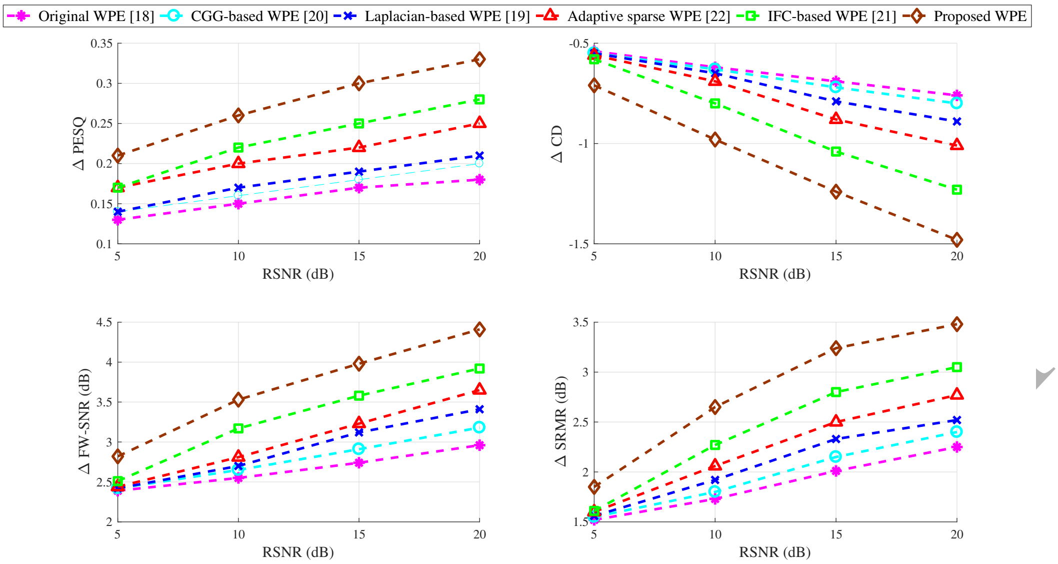

Fig. 11: Performance metrics obtained by using different WPE-based methods using the recorded time-invariant RIRs.

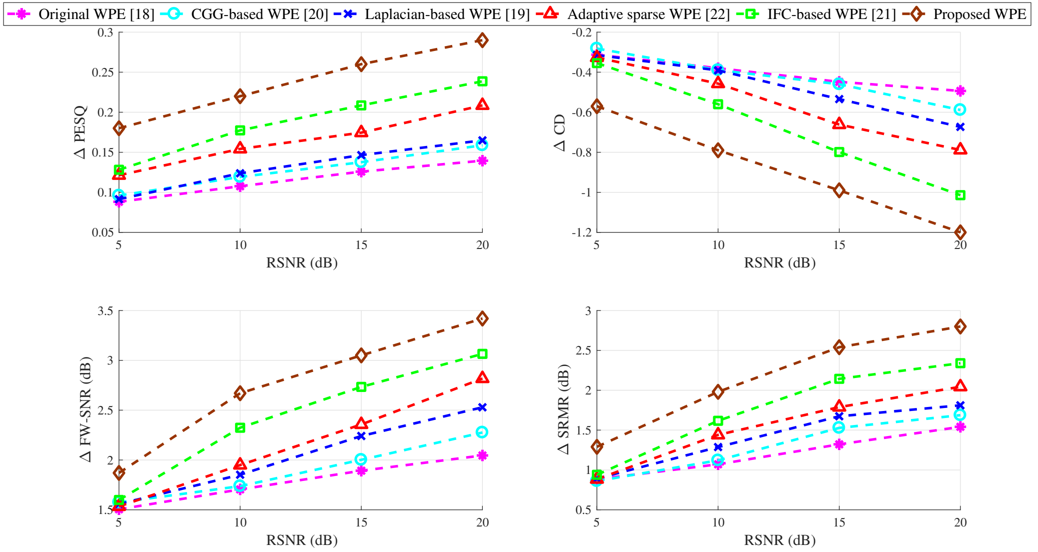

Fig. 12: Performance metrics obtained by using different WPE-based methods using the recorded time-varying RIRs.

Finally, in order to investigate the computational costs involved in the implementation of each dereverberation method, we evaluate the computational efficiency of different methods in terms of the real-time factor (RTF), as in [18]. The RTF can be defined as the ratio of the processing time required for the dereverberation to the time duration of the observed speech. The dereverberation methods were all implemented with MATLAB, and their processing time was measured on a Windows computer with an Intel(R) Core(TM) i5-2320 CPU @ 3.00GHz 3.30GHz with 8.00 GB of RAM. The resultant RTFs averaged over all test utterances with different values of T_{60dB} have been shown in Fig. 14. As observed, while the RTF for the original WPE in [18] and that for the CGG-based WPE in [20] are almost equal, the RTF for the Laplacian-based WPE in [19] is much higher[7]. Correspondingly, it is viewed that the computational burden of the three other methods, namely, the adaptive sparse WPE [22], the IFC-based WPE [21] as well as the proposed WPE in Section 4 are all in the same range with some increase w.r.t. the former two methods. Regarding the proposed approach, taking advantage of the FFT in solving the linear system in (23), it was found that most of the computational effort lies in the TLS-based WPE used to obtain the initial prediction weights as well as the interpolation scheme used to obtain the TV-AR model coefficients.

[7] The main reason for this is the lack of any closed-form solution for the Laplacian-based WPE and the numerical calculation of its corresponding reverberation prediction weights.

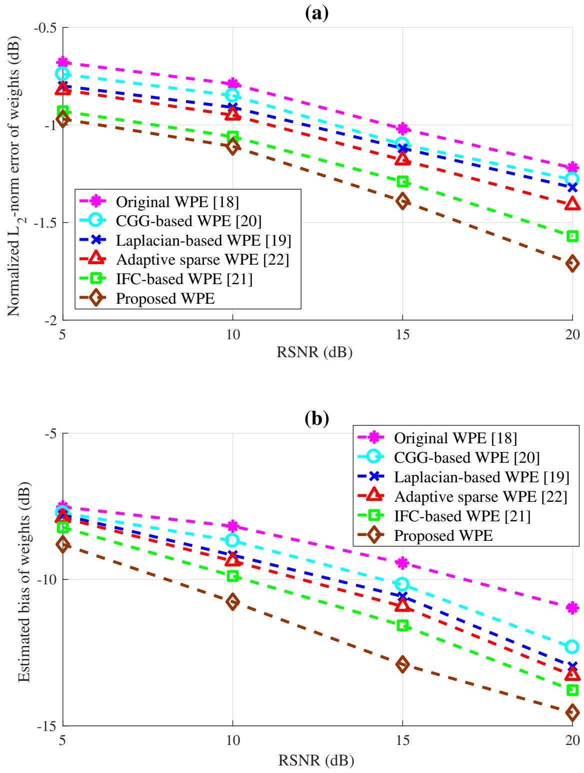

Fig. 13: Error analysis of the prediction weights for different methods w.r.t. the asymptotic weights, (a): Normalized error by (28), (b): Bias given by Jackknife resampling method.

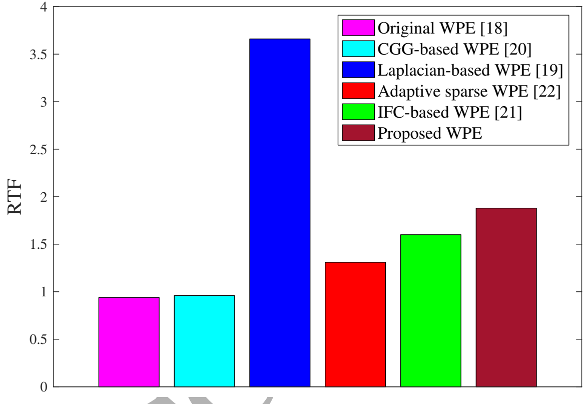

Fig. 14: RTF for each dereverberation method averaged over all test data sets.

6. CONCLUDING REMARKS

We presented a novel approach for blind speech dereverberation in noisy and time-varying environments with no knowledge of the room acoustics or the speech source. We based our approach on the conventional WPE method, yet, since the traditional time-invariant AR model is not realistic in practice, we employed a TV-AR model in order to deal with the variable nature of the speech source as well as the room acoustics. By using the TLS technique to obtain the reverberation prediction weights, $\mathbf{G}$, properly updating $\mathbf{G}$ over time frames and interpolating the resulting $\mathbf{G}$, our multi-folded approach is able to provide superior dereverberation as well as a more robust performance w.r.t. the background noise, particularly under challenging time-varying conditions. The comprehensive performance evaluation presented in both time-invariant and time-varying experiments confirms the advantage of the proposed approach.

Considering the future work within this field, the following directions are of interest:

- Investigation of various TV-AR models (along with methods to estimate their parameters) and integration of them into the WPE method.
- Study of suitable basis functions (other than the DFT used in this work) to form the TV-AR model as in (20)
- Making use of piece-wise AR models (in contrast with TV-AR models) with less complexity in order to handle slowly/moderately changing time-variant RIRs[8]
- Integration of machine learning techniques into the WPE method so that the training (batch) utterances can be different from the testing speech samples.
- Working towards the combination of an initial STFT-domain noise reduction technique with the WPE method to perform joint dereverberation and noise suppression.

7. REFERENCES

[1] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*, Prentice Hall Press, Upper Saddle River, NJ, USA, 2010.

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

[3] 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 Mag.*, vol. 29, no. 6, pp. 114–126, Nov 2012.

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

[5] Y. Huang, J. Benesty, and J. Chen, “A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant

[8] Note that the TV-AR model are generally suitable for non-stationary channels whereas the piece-wise AR models are well suited to piecewise stationary channels.

environment," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 882–895, Sept 2005.

[6] K. Furuya and A. Kataoka, "Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1579–1591, July 2007.

[7] J. M. Yang and H. G. Kang, "Online speech dereverberation algorithm based on adaptive multichannel linear prediction," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 608–619, March 2014.

[8] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multi-channel equalization," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 680–693, April 2016.

[9] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug 2001.

[10] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1529–1539, July 2007.

[11] X. Bao and J. Zhu, "An improved method for late-reverberant suppression based on statistical model," *Speech Communication*, vol. 55, no. 9, pp. 932–940, 2013.

[12] M. Parchami, W. P. Zhu, and B. Champagne, "Model-based estimation of late reverberant spectral variance using modified weighted prediction error method," *Speech Communication*, vol. 92, pp. 100–113, 2017.

[13] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," *Advances in Neural Information Processing Systems*, vol. 13, pp. 758–764, 2001.

[14] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 231–246, Feb 2009.

[15] T. Nakatani, B. H. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "Speech dereverberation based on maximum-likelihood estimation with time-varying Gaussian source model," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1512–1527, Nov 2008.

[16] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, May 2009.

[17] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. 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 (ICASSP)*, March 2008, pp. 85–88, Las Vegas, USA.

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

[19] A. Jukić 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 (ICASSP)*, May 2014, pp. 5172–5176, Florence, Italy.

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

[21] M. Parchami, W. P. Zhu, and B. Champagne, "Speech dereverberation using weighted prediction error with correlated inter-frame speech components," *Speech Communication*, vol. 87, pp. 49–57, 2017.

[22] A. Jukić, T. van Waterschoot, and S. Doclo, "Adaptive speech dereverberation using constrained sparse multichannel linear prediction," *IEEE Signal Processing Letters*, vol. 24, no. 1, pp. 101–105, Jan 2017.

[23] M. G. Hall, A. V. Oppenheim, and A. S. Willsky, "Time-varying parametric modeling of speech," *Signal Processing*, vol. 5, no. 3, pp. 267–285, 1983.

[24] A. Wiesel, O. Bibi, and A. Globerson, "Time varying autoregressive moving average models for covariance estimation," *IEEE Trans. on Signal Processing*, vol. 61, no. 11, pp. 2791–2801, June 2013.

[25] D. Rudoy, T. F. Quatieri, and P. J. Wolfe, "Time-varying autoregressions in speech: Detection theory and applications," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 977–989, May 2011.

[26] C. Sodsri, *Time-Varying Autoregressive Modelling for Nonstationary Acoustic Signal and Its Frequency Analysis*, Ph.D. thesis, The Pennsylvania State University, 2003.

[27] Y. I. Abramovich, N. K. Spencer, and M. D. E. Turley, "Order estimation and discrimination between stationary and time-varying (TVAR) autoregressive models," *IEEE Trans. on Signal Processing*, vol. 55, no. 6, pp. 2861–2876, June 2007.

[28] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 3, pp. 173–185, Mar 2002.

[29] T. Hsiao, "Identification of time-varying autoregressive systems using maximum *a posteriori* estimation," *IEEE Trans. on Signal Processing*, vol. 56, no. 8, pp. 3497–3509, Aug 2008.

[30] S.V. Huffel and J. Vandewalle, *The total least squares problem: Computational aspects and analysis*, Frontiers in applied mathematics, SIAM, Philadelphia, 1991.

[31] M. Clausen, A. Dress, J. Grabmeier, and M. Karpinski, "On zero-testing and interpolation of k-sparse multivariate polynomials over finite fields," *Theoretical Computer Science*, vol. 84, no. 2, pp. 151–164, 1991.

[32] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*, Cambridge University Press, 2012.

[33] H. Rauhut and R. Ward, "Sparse Legendre expansions via ℓ_1-minimization," *Journal of Approximation Theory*, vol. 164, no. 5, pp. 517–533, 2012.

[34] J. Peng, J. Hampton, and A. Doostan, "A weighted ℓ_1-minimization approach for sparse polynomial chaos expansions," *Journal of Computational Physics*, vol. 267, pp. 92–111, 2014.

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

[36] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "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)*, Oct 2013, pp. 1–4, New Paltz, NY, USA.

[37] "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.

[38] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan 2008.

[39] T. H. Falk, C. Zheng, and W. Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, Sept 2010.

[40] Inc. CVX Research, "CVX: Matlab Software for Disciplined Convex Programming, version 2.0," Aug 2012, Available at http://cvxr.com/cvx, last accessed on May 2016.

[41] E. A. Lehmann, "Image-source method: Matlab code implementation," available at http://www.eric-lehmann.com/, last accessed on June 2018.

[42] A. Varga and H. J. M. Steeneken, "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 Communication*, vol. 12, no. 3, pp. 247–251, Jul 1993.

[43] M. Brookes, *VoiceBOX: Speech Processing Toolbox for MATLAB*, 2009, Available at http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox, last accessed on June 2018.

[44] "Recommendation P.56: Objective measurement of active speech level," 1993, ITU-T.

[45] S. Sahinler, "Bootstrap and Jackknife resampling algorithms for estimation of regression parameters," *Journal of Applied Quantitative Methods*, pp. 188–199, 2007.

[46] SimData: dev and eval sets based on WSJCAM0, *REVERB Challenge*, 2013, Available at http://reverb2014.dereverberation.com/download, last accessed on June 2018.

[47] B. Schwarz, "RevDyn Speech Database," Speech and Acoustic Lab of the Faculty of Engineering at Bar Ilan University, Available at http://www.eng.biu.ac.il/schwarb/speech-databases/revdyn-database/, last accessed on June 2018.