Speech Communication 92 (2017) 100–113

Contents lists available at ScienceDirect

Speech Communication

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

Model-based estimation of late reverberant spectral variance using modified weighted prediction error method

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

[a] Department of Electrical and Computer Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal (H3G 1M8), Quebec, Canada
[b] Department of Electrical and Computer Engineering, McGill University, 3480 University Street, Montreal (H3A 0E9), Quebec, Canada

ARTICLE INFO

Article history:
Received 11 November 2016
Revised 5 May 2017
Accepted 21 June 2017
Available online 23 June 2017

Keywords:
Reverberation suppression
Late reverberant spectral variance (LRSV)
Room acoustics
Short-time Fourier transform (STFT)

ABSTRACT

In this paper, we propose a new approach to estimate the late reverberant spectral variance (LRSV) for speech dereverberation in the short-time Fourier transform (STFT) domain. Our approach uses a model-based scheme involving the estimation of a smoothing (shape) parameter and the reverberant-only component of speech. We propose to obtain the shape parameter by using estimates of the spectral variances of the direct-path and reverberant-only components of the speech, which in turn, can be calculated by smoothing coarse estimates of these two components. Furthermore, an accurate estimate of the reverberant-only component is obtained by means of a moving average scheme. In order to obtain the preliminary estimates of the direct-path and reverberant speech components, we employ a modified version of the weighted prediction error (WPE) method. In contrast to the original WPE method, the suggested modification is implemented for shorter processing blocks, each consisting of a number of STFT frames. This block-wise procedure allows for adaptation to moderate changes in environment and makes the proposed approach also suitable for time-varying acoustic scenarios. Performance evaluations with respect to previous LRSV estimation methods demonstrate the superiority of the proposed approach in both time-invariant and time-variant reverberant environments.

1. Introduction

Speech signals captured within an enclosure by a distant microphone are subject to reflections from the surrounding surfaces (walls, ceiling, etc.) and other objects within the environment. This phenomenon, often known as reverberation, can deteriorate the perceived quality/intelligibility of the desired speech signals, and also degrades to a large extent the performance of speech processing systems such as hearing aids, hands-free teleconferencing, source separation and localization, and automatic speech recognition systems (Naylor and Gaubitch, 2010; Yoshioka et al., 2012). Therefore, efficient suppression of reverberation in real world acoustic environments is highly required for these applications.

During the past two decades, numerous single- and multi-microphone dereverberation methods have been developed. In the latter case, the most conventional approaches exploit beamforming techniques to coherently combine the dominant early arrivals, as in e.g., Gannot et al. (2001) and Warsitz and Haeb-Umbach (2007). However, unless a rather large number of microphones is employed, the dereverberation performance of beamforming methods is strictly limited in general (Naylor and Gaubitch, 2010). Many other dereverberation approaches estimate the anechoic (clean) speech by processing observations with inverse filters that can be either calculated using the available room impulse responses (RIRs) or estimated from the reverberant observations (Furuya and Kataoka, 2007; Kumar and Stern, 2010). Even though perfect acoustic equalization is possible in theory if the exact RIR is known, in a realistic acoustic environment, due to the long length and irregular nature of the RIR, its precise estimation is often quite challenging, if not impossible (Habets, 2007). Furthermore, real world RIRs are comprised of several thousands of coefficients in the time or frequency domain, and therefore, their estimation can be a huge task. In this regard, a major stream of research has been focused on the

Abbreviations: (BCI), Blind channel identification; (CD), Cepstrum distance; (DD), Decision-directed; (DRR), Direct-to-reverberant ratio; (EM), Expectation-maximization; (FW-SNR), Frequency-weighted segmental SNR; (ISM), Image source method; (LRSV), Late reverberant spectral variance; (LPC), Linear prediction coefficients; (MA), Moving average; (MMSE), Minimum mean-square 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.2017.06.005
0167-6393/

List of Symbols

Symbol	Definition
n, m	Time index
$x(n)$	Observed reverberant speech
$s(n)$	Anechoic speech
$h_n(m)$	Time-varying RIR
L_h	Length of RIR (samples)
$\hat{s}(n)$	Estimate of anechoic speech
N_e	Length of the early part of RIR (samples)
$h_{E_n}(m)$	Early part of RIR
$h_{L_n}(m)$	Late part of RIR
f_s	Sampling frequency (Hz)
T_{early}	Time length of early RIR (s)
$x_E(n)$	Early part of observed speech
$x_L(n)$	Late part of observed speech
k	Frequency bin index
K	Number of frequency bins (frame length)
l	Time frame index
$w(n)$	STFT analysis-synthesis window
P	STFT frame advance (samples)
$X(k, l)$	STFT of observed speech
$X_E(k, l)$	STFT of early speech
$X_L(k, l)$	STFT of late speech
$\hat{X}_E(k, l)$	Estimate of the STFT of early speech
$\hat{X}_L(k, l)$	Estimate of the STFT of late speech
$G(k, l)$	Gain function of SE
$\zeta(k, l)$	*A priori* signal-to-reverberant ratio
$\eta(k, l)$	*A posteriori* signal-to-reverberant ratio
$\sigma^2_{X_E}$	Spectral variance of early speech
$\sigma^2_{X_L}$	Spectral variance of late speech
$b(m)$	Peak of exponentially decaying RIR, $h(m)$
$T_{60dB}(k)$	60 dB reverberation time (s)
α	$3\log 10/(f_s T_{60dB}(k))$
N_E	N_e/P
$H(k, l)$	RIR model in the STFT domain
$B_D(k)$	Peak of the direct-path component of $H(k, l)$
$B_R(k, l)$	Peak of the reverberant component of $H(k, l)$
$\hat{\sigma}^2_X(k, l)$	Estimate of the spectral variance of entire observed speech
$\hat{\sigma}^2_{X_R}(k, l)$	Estimate of the spectral variance of reverberant-only speech
$\hat{\sigma}^2_{X_L}(k, l)$	Estimate of the spectral variance of late reverberant speech
β	Smoothing parameter in the estimation of $\sigma^2_X(k, l)$
$\kappa(k, l)$	Shape parameter in the estimation of $\sigma^2_{X_R}(k, l)$
$\hat{X}_R(k, l)$	Estimate of reverberant-only speech $X_R(k, l)$
$\sigma^2_{B_D}(k)$	Variance of $B_D(k)$
$\sigma^2_{B_R}(k)$	Variance of $B_R(k, l)$
DRR(k)	Direct-to-reverberant ratio
$\hat{\sigma}^2_{X_D}(k, l)$	Estimate of the spectral variance of direct-path speech
$\hat{\sigma}^2_{X_R}(k, l)$	Estimate of the spectral variance of reverberant-only speech
$\hat{\mathcal{X}}_D(k, l)$	Coarse estimate of direct-path speech
$\hat{\mathcal{X}}_R(k, l)$	Coarse estimate of reverberant-only speech
γ_1, γ_2	Smoothing parameters in the estimation of $\hat{\sigma}^2_{\mathcal{X}_D}(k, l)$ and $\hat{\sigma}^2_{\mathcal{X}_R}(k, l)$
λ	Processing block index (samples)
Δ	Length of processing blocks (samples)
M	Number of speech STFT frames per processing block
d	Number of direct-path speech terms in the WPE method
I	Regression (linear prediction) length in the WPE method
γ	Smoothing parameter for speech variance in the WPE method
ϵ	Flooring value on speech variance in the WPE method
$\mathbf{X}(k, l-d)$	Regression vector in the WPE method
j	Iteration index in the WPE method
J	Number of iterations in the WPE method
$\mathbf{g}_{\lambda_j}(k)$	Reverberation prediction weights in the WPE method
Q	Order of the linear prediction (MA-based) method to estimate $X_L(k, l)$ and $X_R(k, l)$
q	Index of the MA model terms
$X_{de}(k, l)$	Dereverberated speech
$c_q(k, \lambda)$	MA model coefficients
δ	Delay value in the MA model
B	Bias correction factor in the MA model
ρ, ρ'	Thresholds in the criterion for the convergence of $\mathbf{g}_{\lambda_j}(k)$
$F_j(\mathbf{g}_j)$	Cost function used for the estimation of $\mathbf{g}_{\lambda_j}(k)$
$H_j(k, \lambda)$	Thresholded term in the convergence criterion for $\mathbf{g}_{\lambda_j}(k)$
$\mathbf{g}'_\lambda(k)$	Ultimate estimate of reverberation prediction weights (after smoothing)
μ	Smoothing parameter used in the recursive smoothing of $\mathbf{g}'_\lambda(k)$
$e(\Delta)$	Normalized error in the estimation of LRSV
$\mathrm{Err}_{\mathrm{seg}}$	Mean segmental error

use of so-called blind channel identification (BCI) techniques for dereverberation (Huang et al., 2005).

Another important category of reverberation suppression methods includes model-based statistical approaches that target an optimal estimation of the dereverberated speech. In Yoshioka et al. (2009), the parameters of an all-pole model for speech and reverberation are iteratively determined by maximizing the likelihood function of the model parameters through an expectation-maximization (EM) approach. Subsequently, a minimum mean-square error (MMSE) estimator is derived that yields the enhanced speech. As an alternative, the time-varying nature of the speech signal and the multi-channel linear prediction (MCLP) model of reverberation can be exploited for efficient dereverberation, although the implementation of such methods in the time domain is computationally costly (Kinoshita et al., 2009). To overcome this problem and to achieve higher quality in dereverberation, in Nakatani et al. (2008); 2010), it is proposed to implement the MCLP approach 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 speech utterances. However, one of the drawbacks of this method is that it requires at least a few seconds of the observed speech utterance in order to ensure the convergence of the reverberation prediction coefficients. Additionally, the RIR should remain constant during the estimation and dereverberation processes.

Spectral enhancement (SE) methods based on a gain function, originally developed for the purpose of noise reduction, have also

been modified and used for dereverberation (Habets, 2007). The major advantage of SE methods over the aforementioned techniques is their simplicity of implementation in the STFT domain and low computational complexity. In essence, the SE-based dereverberation aims at the suppression of late reverberation, which is the major cause for the deterioration of the speech quality in many scenarios. Assuming that early and late reverberations are independent and under the phase equivalence of the reverberant and anechoic speech, these methods can be employed for late reverberation suppression by estimating the late reverberant spectral variance (LRSV) and using it in place of the noise spectral variance (Habets, 2007). Therefore, the main challenge in reverberation suppression using SE is to estimate the LRSV blindly from reverberant speech observations.

As originally suggested by Lebart et al. in Lebart et al. (2001), the late reverberation can be treated as an additive disturbance. Therein, through a statistical modeling of the RIR, an estimator of the LRSV is derived and used in a spectral subtraction rule. In the same line of work, several estimators of the LRSV have been proposed in the past decade. Since the LRSV estimator in Lebart et al. (2001) is based on a time domain model of the RIR and also assumes that the source-to-microphone distance is larger than a critical distance,[1] Habets developed in Habets et al. (2009) a new LRSV estimator that overcomes these deficiencies. Therein, a statistical RIR model in the STFT domain is proposed and used to derive an extension of the Lebart's LRSV estimator that takes into account the energy contribution of the direct path and reverberant parts of speech. This statistical RIR model only depends on the reverberation time, which is generally almost constant over time. However, similar to Lebart's method, the LRSV estimator in Habets et al. (2009) assumes a fixed RIR, i.e. a time-invariant environment, and also requires the *prior* knowledge of the RIR statistics or the direct-to-reverberant ratio (DRR) parameter. Consequently, this method cannot be implemented blindly, i.e. by processing only the input reverberant speech.

In Erkelens and Heusdens (2010), therefore, an LRSV estimator that is based on the correlation between the reverberant and anechoic speech has been proposed, in contrast to the previous model-based LRSV estimator in Habets et al. (2009). This new LRSV estimator requires no knowledge of the RIR model parameters such as the reverberation time or DRR, and outperforms the previous methods. However, the method in Erkelens and Heusdens (2010) can only track very slow changes in the RIR and underestimates the LRSV in case of time-varying RIRs. Therefore, it is suggested in Erkelens and Heusdens (2010) to use a model-based LRSV estimation for the general case of time-varying RIRs. While it is shown therein that this scheme is advantageous in slowly changing environments, the amount of data needed for the blind estimation of the required shape parameter is on the order of several seconds. This does not allow the developed scheme in Erkelens and Heusdens (2010) to be adapted to real world changing environments.

Along the same direction, a few recursive smoothing schemes for LRSV estimation have been suggested in the recent literature, such as the one in Bao and Zhu (2013). Therein, since the smoothing (or the so-called shape) parameter is affected by the estimation errors of the LRSV, it is suggested therein to use more than one term of the past spectral variances of the reverberant speech in the recursive smoothing scheme used for the calculation of the shape parameter. However, only minor improvements can be observed by using the latter method with respect to the previous schemes in Habets et al. (2009) and Erkelens and Heusdens (2010). In summary, it is concluded that despite the existence of a few major schemes for the estimation of the LRSV, blind estimation of this parameter, particularly in changing acoustic environments, remains a challenging problem.

In this work, we present a new approach for the estimation of the LRSV that relies on the statistical model-based method in Habets et al. (2009). Our approach mainly targets the task of speech enhancement in highly reverberant environment, even though it can be further extended and employed for other related tasks such as speech recognition. The new approach uses a recursive smoothing scheme which requires the proper selection of the underlying shape parameter as well as an accurate estimate of the reverberant-only speech component. To approximate the optimal shape parameter, we employ the spectral variances of the direct-path and reverberant-only speech components, which in turn, can be estimated by smoothing coarse estimates of these components. Further, to obtain an accurate estimate of the reverberant-only speech, we employ a moving average (MA) scheme which requires a coarse estimate of the direct-path speech component. To obtain the coarse estimates of the direct-path and reverberant-only components, we take advantage of the WPE dereverberation method. Yet, in contrast to the original WPE method, which requires the entire set of speech observations to estimate the underlying reverberation prediction weights, we implement the WPE method in an incremental fashion, where the observed speech is processed block by block. This makes the overall proposed LRSV estimation approach suitable for changing environments where the reverberation prediction weights have to be adapted over time.

This paper is organized as follows. In Section 2, a brief overview of late reverberation suppression using the SE method and the estimation of LRSV is presented. The proposed approach for LRSV estimation is developed in Section 3. Section 4 is devoted to performance evaluation via experimentation and Section 5 concludes the paper.

2. Background

In this section, we first present a brief background on late reverberation suppression based on a gain function. Next, we review the model-based method for the estimation of LRSV, which is the most critical component in the calculation of gain functions.

2.1. Reverberation suppression using a gain function

In an acoustic environment, the captured reverberant signal by a microphone, $x(n)$, with $n \in \{0, 1, \cdots\}$ as the discrete-time index, can be modeled in the time domain as the convolution of the anechoic speech, $s(n)$, with the causal time-varying RIR, $h_n(m)$, where $m \in \{0, 1, \cdots, L_h\}$ denotes the sample index and L_h is the length of the RIR (Naylor and Gaubitch, 2010):

$$x(n) = \sum_{m=0}^{L_h-1} h_n(m)s(n-m) \tag{1}$$

The ultimate goal of dereverberation is to obtain an estimate of the anechoic speech signal denoted as $\hat{s}(n)$, using the observation signal $x(n)$. The problem of interest is termed as blind dereverberation, since neither the speech signal $s(n)$ (and its characteristics) nor the acoustic RIR $h_n(m)$ is available. Since our aim is to suppress the late reverberant speech[2], we divide the RIR in (1) into the early and late parts as

$$h_n(m) = \begin{cases} h_{E_n}(m), & 0 \le m < N_e \\ h_{L_n}(m), & N_e \le m < L_h \\ 0, & \text{Otherwise} \end{cases} \tag{2}$$

[1] In smaller distances, the LRSV estimator in Lebart et al. (2001) largely overestimates the LRSV.

[2] It should be noted that in most applications of speech enhancement in reverberant environments, the main cause of degradation in the quality/intelligibility of speech is the late reverberant component with the early component even improving noise-related measures such as the SNR (Habets, 2007).

with $h_{E_n}(m)$ and $h_{L_n}(m)$ as the early and late parts of the RIR respectively and N_e as the length of early reflections. In practice, N_e can be obtained as $f_s T_{early}$ with f_s being the sampling frequency in Hz and T_{early} as the early speech duration ranging from 40 to 80 ms (Naylor and Gaubitch, 2010). Inserting (2) into (1) results in

$$x(n) = \underbrace{\sum_{m=0}^{N_e-1} h_{E_n}(m)s(n-m)}_{x_E(n)} + \underbrace{\sum_{m=N_e}^{L_h-1} h_{L_n}(m)s(n-m)}_{x_L(n)} \tag{3}$$

where $x_E(n)$ and $x_L(n)$ can be respectively referred to as the early and late reverberant components of speech. In the STFT domain, by denoting the frequency bin and time frame indices respectively as $k \in \{0, 1, \cdots, K-1\}$ with K as the number of total frequency bins and $l \in \mathbb{N}$, it follows that

$$X(k,l) = \sum_{n=0}^{K-1} x(n+lP)w(n)e^{-j\frac{2\pi k}{K}n} = X_E(k,l) + X_L(k,l) \tag{4}$$

where $w(n)$ is the analysis window, P is the STFT frame advance (hop size), and $X_E(k, l)$ and $X_L(k, l)$ are respectively the early and late reverberant components in the STFT domain.

The goal of late reverberation suppression by means of SE is to obtain an estimate of the early reverberant component, $X_E(k, l)$. This was originally accomplished by Lebart et al. (2001) where the conventional spectral subtraction rule, initially developed for additive noise reduction, is applied on the reverberant observation through a multiplicative gain function, namely,

$$\hat{X}_E(k,l) = G(k,l)X(k,l) \tag{5}$$

with $\hat{X}_E(k, l)$ and $G(k, l)$ respectively as the estimated early reverberant speech and spectral gain function. Various expressions for $G(k, l)$ can be found from the noise reduction literature, e.g., those employed in Lebart et al. (2001). In turn, this gain function generally depends on two important parameters, which in the context of late reverberation suppression, are

$$\zeta(k,l) = \frac{\sigma^2_{X_E}(k,l)}{\sigma^2_{X_L}(k,l)}, \quad \eta(k,l) = \frac{|X(k,l)|^2}{\sigma^2_{X_L}(k,l)} \tag{6}$$

where the two parameters $\sigma^2_{X_E}(k,l) = E\{|X_E(k, l)|^2\}$ and $\sigma^2_{X_L}(k,l) = E\{|X_L(k, l)|^2\}$ are respectively the spectral variances of the early and late reverberant components. Borrowing from the noise reduction context, $\zeta(k, l)$ can be estimated through the conventional decision-directed (DD) approach (Ephraim and Malah, 1984), using an estimate of $\sigma^2_{X_L}(k,l)$. Within this framework, the estimation of $\sigma^2_{X_L}(k,l)$, i.e. the so-called LRSV, due to its high influence on the overall performance of the SE method, has received considerable attention in the recent literature and is therefore the main focus of this work.

2.2. Model-based method for LRSV estimation

Polack (1988) originally modeled an RIR in the time domain as

$$h(m) = b(m)e^{-\alpha m} \tag{7}$$

where $b(m)$ is a zero-mean white Gaussian random process, and α is defined as $3\log 10/(f_s T_{60dB})$ with f_s as the sampling frequency in Hz and T_{60dB} as the 60dB reverberation time in seconds. Based on this model, Lebart et al. (2001) derived the following estimator for the LRSV

$$\hat{\sigma}^2_{X_L}(k,l) = e^{-2\alpha P N_E}\sigma^2_X(k, l-N_E) \tag{8}$$

with $\sigma^2_X(k,l) = E\{|X(k, l)|^2\}$ denoting the spectral variance of the observation and $N_E = N_e/P$. It should be noted that the independence of the early and late reverberant components has been assumed to derive (8) and also in other prominent LRSV estimators.

In Habets et al. (2009), Habets suggests the following statistical RIR model in the STFT domain

$$H(k,l) = \begin{cases} B_D(k), & l = 0 \\ B_R(k,l)e^{-\alpha(k)lP}, & l \geq 1 \end{cases} \tag{9}$$

where $B_D(k)$ and $B_R(k, l)$ are zero-mean mutually independent and identically distributed (i.i.d.) Gaussian random processes corresponding respectively to the direct-path and reverberant components of the RIR. Note that in this model, $\alpha(k)$ has been defined as $3\log 10/(f_s T_{60dB}(k))$ with $T_{60dB}(k)$ considered as a frequency-dependent parameter. Based on this model, a recursive scheme for the LRSV estimator is derived in Habets et al. (2009) as given below

$$\hat{\sigma}^2_X(k,l) = [1-\beta]\hat{\sigma}^2_X(k,l-1) + \beta|X(k,l)|^2 \tag{10a}$$

$$\hat{\sigma}^2_{X_R}(k,l) = [1-\kappa(k)]e^{-2\alpha(k)P}\hat{\sigma}^2_{X_R}(k,l-1) \tag{10b}$$

$$+\kappa(k)e^{-2\alpha(k)P}\hat{\sigma}^2_X(k,l-1)$$

$$\hat{\sigma}^2_{X_L}(k,l) = e^{-2\alpha(k)P(N_E-1)}\hat{\sigma}^2_{X_R}(k,l-N_E+1) \tag{10c}$$

with β=0.15 as a fixed smoothing parameter and $\kappa(k)$ the shape parameter used to estimate the reverberant spectral variance $\sigma^2_{X_R}(k,l)$. Herein, the reverberant component $X_R(k, l)$ is in fact the entire reverberant speech $X(k, l)$ except the direct-path (first) term.

Since $\sigma^2_{X_R}(k,l)$ should exclude the direct-path speech component in order to avoid distorting this component by the underlying spectral suppression rule, the selection of the shape parameter $\kappa(k)$ is of high importance. In Habets et al. (2009), it is proved that the optimal value of this parameter is in fact the ratio of the variance of $B_R(k, l)$ to that of $B_D(k)$, which can be obtained by

$$\kappa(k) = \frac{\sigma^2_{B_R}(k)}{\sigma^2_{B_D}(k)} = \frac{e^{2\alpha(k)P}-1}{\mathrm{DRR}(k)} \tag{11}$$

where DRR(k) is the so-called direct-to-reverberant ratio defined as the ratio of the energy of the direct-path RIR to that of the reverberant RIR. However, the use of (11) poses a number of difficulties. First, DRR(k) has to be estimated beforehand in a blind manner, implying an additional task requiring at least a few seconds of reverberant observations. Secondly, this scheme does not properly suit the case of a changing environment (RIR). Thirdly, as observed from (10), the estimation of the reverberant spectral variance $\sigma^2_{X_R}(k,l)$ is performed by recursively smoothing the entire reverberant observation $X(k, l)$, and therefore, the estimated $\hat{\sigma}^2_{X_R}(k,l)$ includes the direct-path component of speech as well.

In the following section, we propose a new scheme for the estimation of the LRSV, which is suitable for moderately changing environments. Our scheme takes advantage of a linear prediction-based dereverberation method in eliminating the direct-path component when estimating the reverberant spectral variance $\sigma^2_{X_R}(k,l)$.

3. Proposed LRSV estimator

Although our approach for estimating the LRSV is based on the scheme in Habets et al. (2009), as discussed in the previous section, we target time-varying acoustic environments where the RIR cannot be assumed constant over a period of a few seconds. In this respect, as opposed to (10a) and (10b), we use the following scheme for the estimation of the reverberant-only spectral variance:

$$\hat{\sigma}^2_{X_R}(k,l) = [1-\kappa(k,l)]\hat{\sigma}^2_{X_R}(k,l-1) + \kappa(k,l)|\hat{X}_R(k,l)|^2 \tag{12}$$

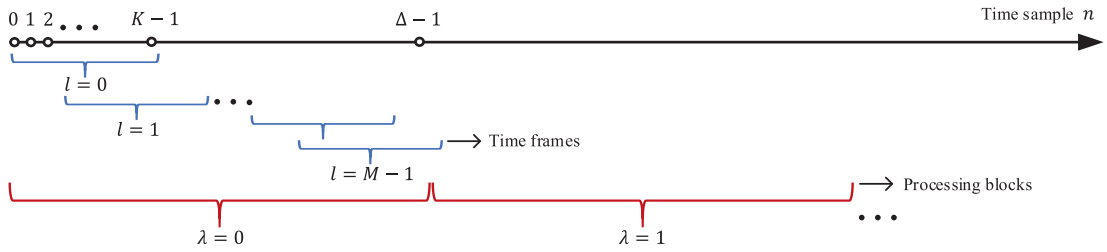

Fig. 1. An illustration of the STFT frames and the processing blocks over speech time samples.

and then use (10c) to obtain the LRSV, $\sigma^2_{X_L}(k,l)$. As compared to (10b), a new time and frequency dependent scheme for the shape parameter $\kappa(k,\, l)$ is proposed, which fits properly the case of a time-varying RIR. In addition, rather than estimating the reverberant-only spectral variance $\sigma^2_{X_R}(k,l)$ by smoothing $|X(k,\, l)|^2$, we will exploit an estimate of the reverberant-only speech, $\hat{X}_R(k,l)$, which excludes the direct-path component. This, to a large extent, helps avoiding the leakage of the direct-path speech into the estimated LRSV. In Sections 3.1 and 3.2 below, we will respectively present the proposed schemes for the shape parameter $\kappa(k,\, l)$ and the estimation of the reverberant-only component $\hat{X}_R(k,l)$.

3.1. Suggested scheme for the shape parameter

Based on (11), we propose a new blind scheme to obtain the shape parameter κ. This is achieved by finding a proper estimator for the DRR(k) in (11) as a function of time frame l and frequency bin k. In this regard, we propose to choose the shape parameter by the following

$$\kappa^1(k,l) = \frac{e^{2\alpha(k)P} - 1}{\hat{\sigma}^2_{\mathcal{X}_D}(k,l)/\hat{\sigma}^2_{\mathcal{X}_R}(k,l)}$$
$$\kappa(k,l) = \min\{\max\{\kappa^1(k,l), 0\}, 1\} \tag{13}$$

where $\hat{\sigma}^2_{\mathcal{X}_D}(k,l)$ and $\hat{\sigma}^2_{\mathcal{X}_R}(k,l)$ are the estimate of the spectral variance of the direct-path and that of the reverberant speech, respectively, and the second equation is to ensure that the shape parameter lies in [0, 1]. To estimate the two spectral variances in (13), we use the recursive smoothing method, i.e.,

$$\hat{\sigma}^2_{\mathcal{X}_D}(k,l) = [1-\gamma_1]\hat{\sigma}^2_{\mathcal{X}_D}(k,l-1) + \gamma_1|\hat{\mathcal{X}}_D(k,l)|^2$$
$$\hat{\sigma}^2_{\mathcal{X}_R}(k,l) = [1-\gamma_2]\hat{\sigma}^2_{\mathcal{X}_R}(k,l-1) + \gamma_2|\hat{\mathcal{X}}_R(k,l)|^2 \tag{14}$$

with γ_1 and γ_2 as two fixed smoothing parameters taken to be 0.25, and $\hat{\mathcal{X}}_D(k,l)$ and $\hat{\mathcal{X}}_R(k,l)$ as coarse estimates of the direct-path and reverberant components of speech, respectively. Here, we resort to a linear prediction-based dereverberation method in the STFT domain, namely the WPE method (Nakatani et al., 2010), in order to obtain $\hat{\mathcal{X}}_D(k,l)$ and $\hat{\mathcal{X}}_R(k,l)$. However, the WPE method is in essence a batch processing technique and it requires the preprocessing of the entire speech utterance in order to provide an accurate performance. This is not suitable when dealing with a time-varying acoustic environment, where the RIR is prone to change. Furthermore, a large processing delay is introduced due to the preprocessing step, which is undesirable for real-time processing of the speech. To overcome these obstacles, here, we employ the WPE method for processing blocks of typically 0.5 s long. We then exploit the estimated direct-path and reverberant components obtained from the WPE method for $\hat{\mathcal{X}}_D(k,l)$ and $\hat{\mathcal{X}}_R(k,l)$, respectively, in (14) at each processing block. A schematic of the processing blocks and the corresponding frames over time is shown in Fig. 1. Within this framework, the resulting coarse estimates of the direct-path and reverberant components are precise enough for the suggested scheme for $\kappa(k,\, l)$ in (13) and (14), as will be investigated thoroughly in Section 4.

Now, denoting each processing block by λ and the block length (in samples) by Δ, based on Nakatani et al. (2010), the resulting block-wise WPE method can be summarized as follows:

- At the processing block λ, the observation $X(k,\, l)$ is considered for $l \in \{\lambda M, \lambda M+1, \cdots, \lambda M+M-1\}$ (which are actually M STFT frames). We set the following parameters: the number of direct-path speech terms $d = 1$, the regression (linear prediction) length $I = 15$, the smoothing parameter for speech variance $\gamma = 0.65$ and the flooring value on speech variance $\epsilon = 10^{-3}$. Next, we form the regression vector $\mathbf{X}(k, l-d)$ as below

$$\mathbf{X}(k,l-d) = [X(k,l-d), X(k,l-d-1), \cdots \cdots, X(k,l-d-I+1)]^T \tag{15}$$

- The speech spectral variance $\sigma^2_{\mathcal{X}_D}(k,l)$ is initialized as $\sigma^2_{\mathcal{X}_{D_0}}(k,l) = |X(k,l)|^2$.
- Repeat the following for j from 0 to $J-1$, with J as the number of iterations

$$\mathbf{A}_{\lambda_j}(k) = \sum_l \frac{\mathbf{X}(k,l-d)\mathbf{X}^H(k,l-d)}{\sigma^2_{\mathcal{X}_{D_j}}(k,l)}$$
$$\mathbf{a}_{\lambda_j}(k) = \sum_l \frac{\mathbf{X}(k,l-d)X^*(k,l)}{\sigma^2_{\mathcal{X}_{D_j}}(k,l)} \tag{16}$$

where $l \in \{\lambda M, \lambda M+1, \cdots, \lambda M+M-1\}$, and $\{.\}^H$ and $\{.\}^*$ respectively denote matrix hermitian and complex conjugation.

$$\mathbf{g}_{\lambda_j}(k) = \mathbf{A}^{-1}_{\lambda_j}(k)\ \mathbf{a}_{\lambda_j}(k) \tag{17}$$

$$\mathcal{X}_{R_j}(k,l) = \mathbf{g}^H_{\lambda_j}(k)\,\mathbf{X}(k,l-d)$$
$$\mathcal{X}_{D_j}(k,l) = X(k,l) - \mathcal{X}_{R_j}(k,l) \tag{18}$$

$$\sigma^2_{\mathcal{X}_{D_{j+1}}}(k,l) = [1-\gamma]\sigma^2_{\mathcal{X}_{D_{j+1}}}(k,l-1) + \gamma \max\left\{|\mathcal{X}_{D_j}(k,l)|^2, \epsilon\right\} \tag{19}$$

- The terms $\mathcal{X}_{R_j}(k,l)$ and $\mathcal{X}_{D_j}(k,l)$ at the last iteration are considered as $\hat{\mathcal{X}}_R(k,l)$ and $\hat{\mathcal{X}}_D(k,l)$ in (14).

Note that, contrary to the original WPE method, here the reverberation prediction weights $\mathbf{g}_{\lambda_j}(k)$ are estimated separately for each time block λ. Also, to obtain a smoother speech spectral variance $\sigma^2_{\mathcal{X}_D}(k,l)$, which reasonably enhances the overall performance, a smoothing scheme has been considered for this parameter in (19) rather than its instantaneous estimate used in the original method. In our case, the parameter setting $d = 1$ should be considered so that $\mathcal{X}_{D_j}(k,l)$ in (18) particularly estimates the direct-path component of speech. Even though the WPE method is often implemented for a fixed number of iterations J, we use a more efficient heuristic criterion for the number of iterations, which will be discussed in Section 3.3.

3.2. Estimation of the reverberant component

To obtain a proper estimate of the reverberant-only component of the speech, we modify the correlation-based approach suggested in Erkelens and Heusdens (2010), which was originally proposed to estimate the late reverberant component. This approach models the estimate of the late reverberant speech, $\hat{X}_L(k,l)$, as a weighted sum of Q previous frames of the dereverberated (direct-path) speech, as the following

$$\hat{X}_L(k,l) = \sqrt{B}\sum_{q=0}^{Q-1} c_q(k) X_{de}(k, l-\delta-q) \tag{20}$$

where $X_{de}(k,\ l)$ is the dereverberated speech, δ is a delay (in the order of a few frames) to skip the direct-path and early reverberant components, c_q's are the MA model (prediction) coefficients, $Q = 60$ is the number of MA terms and $B = 1.65$ is a bias correction factor (Erkelens and Heusdens, 2010). Since we here aim at the estimation of the entire reverberant speech including the early and late components, we set $\delta = 1$ in the above to only skip the direct-path component at the current frame and use the direct-path component obtained from the WPE method in Section 3.1 for $X_{de}(k,\ l)$. This results in

$$\hat{X}_R(k,l) = \sqrt{B}\sum_{q=0}^{Q-1} c_q(k,\lambda)\hat{\mathcal{X}}_D(k, l-1-q) \tag{21}$$

where we have used the term $\hat{\mathcal{X}}_D(k,l)$ as an estimate for $X_{de}(k,\ l)$. Also, in a similar fashion to the reverberation prediction weights $\mathbf{g}_\lambda(k)$, we have considered the prediction coefficients $c_q(k,\ \lambda)$ to be updated as a function of the block index λ to account for moderate changes in the environment. Now, what remains is to obtain the prediction coefficients $c_q(k,\ \lambda)$, as required by (21). As in Erkelens and Heusdens (2010), the prediction coefficients can be optimally obtained by minimizing the mean squared error between $X(k,\ l)$ and $c_q(k,\lambda)\hat{\mathcal{X}}_D(k, l-1-q)$, which leads to the following solution

$$\hat{c}_q(k,\lambda) = \frac{E_l\{X(k,l)\hat{\mathcal{X}}_D(k,l-1-q)\}}{E_l\{|\hat{\mathcal{X}}_D(k,l-1-q)|^2\}} \tag{22}$$

where $E_l\{.\}$ denotes the expectation over frames. Even though calculating $E_l\{.\}$ requires long-term time averaging, here, the block processing framework allows to perform the time averaging with enough number of frames. In this sense, denoting the terms in the numerator and denominator of (22) by $E^{(1)}$ and $E^{(2)}$, respectively, we use the following sample averaging

$$\begin{aligned} E^{(1)} &\approx \frac{1}{M}\sum_l X(k,l)\hat{\mathcal{X}}_D(k,l-1-q) \\ E^{(2)} &\approx \frac{1}{M}\sum_l |\hat{\mathcal{X}}_D(k,l-1-q)|^2 \end{aligned} \tag{23}$$

where we let $l \in \{\lambda M, \lambda M+1, \cdots, \lambda M+M-1\}$, i.e., we perform the sample means over the M frames of the processing block, λ. It should be noted that, even though the block-wise implementation of the WPE method, as discussed in Section 3.1, introduces deviations in the prediction weights $\mathbf{g}_\lambda(k)$ from those obtained through the full batch processing, the WPE method still does a good job at isolating the direct-path component from the reverberant one as obtained by (21). Further details regarding the performance of the WPE method-based on block processing will be further discussed in Section 4.

In Fig. 2, a block diagram of the main steps of the proposed approach for LRSV estimation is illustrated. It is observed that the estimates of the direct and reverberant components by the WPE method are used for updating both the shape parameter $\kappa(k,\ l)$ and the reverberant component $\hat{X}_R(k,l)$, as required in the proposed LRSV estimation scheme.

3.3. Implementation of the WPE method

The original WPE method essentially requires batch processing using at least a few seconds of the reverberant observation. In spite of this, we apply the WPE method for processing blocks of 0.5 s, since it is employed only to provide preliminary estimates of the direct-path and reverberant speech components. Furthermore, to make the underlying WPE method suitable for our block processing-based approach, we make a few modifications to the original version of this method. First, as discussed in Section 3.1, a smoothing scheme is added for the estimation of the speech spectral variance $\sigma^2_{\mathcal{X}_D}(k,l)$ in (19). Next, we employ a heuristic criterion for the number of iterations performed in (15)–(19). Conventionally, a fixed or a maximum number of iterations can be employed, or more precisely, the following convergence criterion can be used at the jth iteration (Yoshioka, 2010)

$$\frac{\left\|\mathbf{g}_j(k) - \mathbf{g}_{j-1}(k)\right\|_2}{\left\|\mathbf{g}_{j-1}(k)\right\|_2} < \rho \tag{24}$$

with $\|.\|_2$ denoting the ℓ_2-norm and ρ as a fixed threshold value; The iterations are discarded if the above holds. Here, we suggest a convergence criterion based on a heuristic interpretation of the WPE method in Yoshioka (2010), as follows. The reverberation prediction weights $\mathbf{g}_j(k)$ can actually be derived based on the minimization of the following cost function (Yoshioka, 2010)

$$\begin{aligned} F_j(\mathbf{g}_j) &= \sum_l \frac{\left|X(k,l) - \mathbf{g}_j^H(k)\mathbf{X}(k,l-d)\right|^2}{|\mathcal{X}_{D_{j-1}}(k,l)|^2} \\ &= \sum_l \frac{|\mathcal{X}_{D_j}(k,l)|^2}{|\mathcal{X}_{D_{j-1}}(k,l)|^2} \end{aligned} \tag{25}$$

which in fact penalizes the sparsity of the dereverberated speech in the numerator as compared to the anechoic speech in the denominator. Here, we take advantage of the criterion expressed in (25) to formulate a more efficient convergence criterion than the one in (24) for the reverberation prediction weights at the λ-th processing block, $\mathbf{g}_{\lambda_j}(k)$, as the following

$$H_j(k,\lambda) = \sum_{l=\lambda M}^{\lambda M+M-1} \frac{|\mathcal{X}_{D_j}(k,l)|^2}{|\mathcal{X}_{D_{j-1}}(k,l)|^2} < \rho' \tag{26}$$

where the summation is performed on all frames of the λ-th processing block and the threshold value ρ' is experimentally set to $0.01M$. This choice of the convergence criterion ensures that a certain level of sparsity in the dereverberated speech, as inspired by the cost function in (25), is reached before discarding the iterations. Since the values of $|\mathcal{X}_{D_j}(k,l)|^2$ and $|\mathcal{X}_{D_{j-1}}(k,l)|^2$ may change dramatically in some time-frequency units, we set the maximum allowed number of iterations to 10.

Finally, to smooth the changes of the reverberation prediction weight $\mathbf{g}_\lambda(k)$ across processing blocks, we perform a smoothing scheme on $\mathbf{g}_\lambda(k)$ to obtain its ultimate value, $\mathbf{g}'_\lambda(k)$, as

$$\mathbf{g}'_\lambda(k) = [1-\mu]\mathbf{g}'_{\lambda-1}(k) + \mu\mathbf{g}_\lambda(k) \tag{27}$$

with μ fixed at 0.8, to update the values of $\mathbf{g}'_\lambda(k)$ by using mostly the current processing block.

3.4. Relation between the SE and WPE methods

The SE methods were originally developed based on a gain function for the purpose of noise reduction (Loizou, 2013), and later, they were modified in order to handle the late reverberation suppression problem (Habets, 2007). This group of methods mainly models the additive disturbances (noise or reverberation) by a zero mean complex Gaussian distribution and aims at estimating the

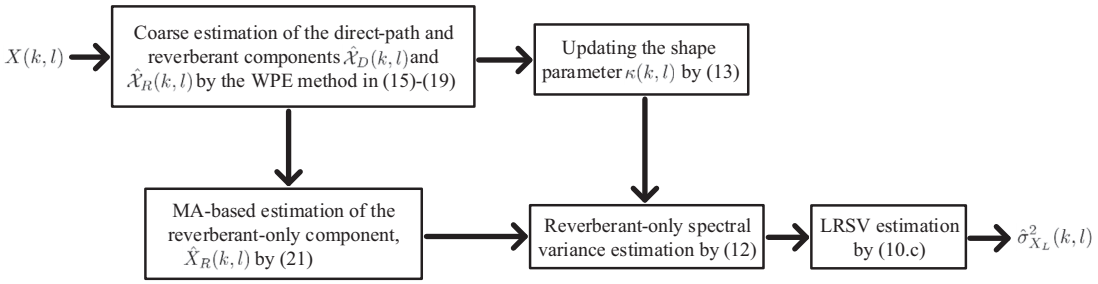

Fig. 2. Block diagram of the proposed algorithm for LRSV estimation.

clean speech by means of a maximum likelihood (ML), maximum *a posteriori* (MAP) or Bayesian MMSE approach. More precisely, it follows that

$$X(k,l) = S(k,l) + V(k,l) \tag{28}$$

with $S(k,l) = A(k,l)e^{j\theta(k,l)}$ as the clean speech and $V(k,l)$ as the sum of additive disturbance terms. Classic SE techniques tackle this problem by maximizing the likelihood, $p(X|\hat{A})$, or the MAP distribution, $p(\hat{A}|X)$, or by minimizing the expectation of a cost function such as $(A-\hat{A})^2$ w.r.t. the speech amplitude estimate $\hat{A}$ (Loizou, 2013). Given that a proper estimate of the disturbance spectral variance (the LRSV in our case), $\sigma_V^2(k,l)$, is at hand, the gain function $G(k,l)$ can be computed and multiplied by the observed speech $X(k,l)$ in order to suppression the disturbance. The WPE method, on the other hand, is a fully blind method which need not any *prior* knowledge of the acoustic environment or disturbance statistics. This method is based on a linear predictive modeling for the disturbance $V(k,l)$, expressed as $\mathbf{g}^H(k)\mathbf{X}(k,l)$ or $\sum_{\ell=1}^{l} g^*(k,\ell)X(k,l-\ell)$. Using the WPE method, one is able to estimate the linear prediction weights $\mathbf{g}(k)$ and thus the clean speech $S(k,l)$ in a fully blind way by means of an ML method, as the following (Nakatani et al., 2010)

$$\hat{\mathbf{g}} = \max_{\mathbf{g}} \; p(S|\mathbf{g}) = \max_{\mathbf{g}} \; p\big((X - \mathbf{g}^H\mathbf{X})|\mathbf{g}\big) \tag{29}$$

According to the linear predictive model used for $V(k,l)$ with I being a rather large number, $V(k,l)$ is in fact a sum of many random terms. This, based on the central limit theorem, implies that the disturbance $V(k,l)$ is assumed to be approximately normally distributed, as is the case with SE methods. Therefore, it can be concluded that both the WPE and SE methods rely on almost the same set of assumptions. The main difference is that, however, the SE methods provide no means of estimating the disturbance spectral variance, necessitating the use of an LRSV estimator when used for dereverberation. On the other hand, the WPE method relies on a set of STFT frames, namely a processing block, in order to obtain the prediction weights $\mathbf{g}(k)$.

It is well-known that the biggest challenge in modifying the SE technique for the task of dereverberation is the estimation of the LRSV (Habets, 2007). In the model-based LRSV estimation problem of interest, we found that the biggest obstacle in using the smoothing scheme of (12) is the lack of proper estimates for the direct-path and reverberant-only speech components. Given that the WPE method can reliably provide preliminary estimates of these two speech components, we employed a modification of this blind method in our approach, where the number of early speech terms d is set to 1, to only account for the direct-path component (first speech term). Yet, in order to make the underlying WPE method compatible with the online estimation of LRSV, we used the block-wise WPE method described in (15)–(19) to have the smallest possible processing delay.

4. Performance evaluation

4.1. Methodology

In this section, we evaluate the performance of the proposed LRSV estimator as compared to a few major LRSV estimation methods for both time-invariant and time-varying RIRs. To this end, anechoic speech utterances including 10 male and 10 female speakers are used from the TIMIT database (Garofolo et al., 1993), the sampling frequency f_s is set to 16 kHz and a 25 ms Hamming window with overlap of 75% is used for the STFT analysis-synthesis. To implement our block processing-based approach, we consider a block length of 0.5 s, resulting in M=80 frames in each processing block.[3] It should be noted that there exists a trade-off in choosing the length of processing blocks, since the shorter the block length the more erroneous the prediction weights $\mathbf{g}_\lambda(k)$ whereas the longer the block length the higher the processing delay and also the slower the adaptation of the estimated LRSV to the changing RIR. With the current choice for the processing block length, considering the computational complexity of the underlying WPE method,[4] the proposed approach seems suitable for real time applications in which the dereverberation algorithm needs to be performed incrementally from the beginning of the captured speech utterance with a small algorithmic delay. To obtain the best performance, T_{early} is chosen to be 62.5 ms, resulting in N_E=10 for our experiments. As for the estimation of the reverberation time T_{60dB}, we use the blind reverberation time estimator in Löllmann et al. (2010) which is capable of estimating T_{60dB} within the allowed processing blocks with low complexity and enough accuracy for our LRSV estimation method. Note that, even for mildly changing environments, the reverberation time T_{60dB} does not often change considerably (Naylor and Gaubitch, 2010). Our approach does not require the estimation of the DRR parameter. As opposed to Erkelens and Heusdens (2010), where Q in (20) was taken as 60 to account for heavy reverberations with T_{60dB}'s of up to 2 s, we choose Q in (21) to be 20 to deal with moderate amounts of reverberation but we increase the bias correction factor to $B = 3.2$.

For the evaluation of the reverberation suppression achieved by using the proposed approach in a spectral suppression rule, we use four performance measures recommended by REVERB Challenge (Kinoshita et al., 2013). These performance metrics 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 and is the one recommended by ITU-T standards for speech quality assessment

[3] Note that M can be calculated by dividing the block length Δ by the STFT hop size P.

[4] This has been studied in detail in (Nakatani et al., 2010) in terms of the real time factor.

(Recommendation P.862, 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 following, we evaluate the relative performance of the proposed LRSV estimator in both time-invariant and time-varying reverberant environments.

4.2. Performance in time-invariant RIRs

In this part, we assess the performance of the proposed approach in comparison with other methods in a scenario where the environment is invariant using both synthesized and recorded RIRs. In case of the recorded RIR, we use the measured RIR from SimData of the REVERB Challenge (Kinoshita et al., 2013), where an 8 channel circular array with diameter of 20 cm was placed in a 3.7 m × 5.5 m acoustic room.[5] The resulting signal was contaminated with additive babble noise from the same database at a global reverberant SNR of 10 dB. Furthermore, in order to verify the performance of the proposed approach in different amounts of reverberation, we use the image source method (ISM) Lehmann to synthesize RIRs with controllable T_{60dB}. In all cases, the anechoic speech is convolved with the RIR to obtain the reverberant speech signal. The geometry of the synthesized reverberant environment with T_{60dB} ranging from 100 ms to 800 ms is shown in Fig. 3. The global SNR is fixed at 15 dB for this experiment.

In case of the time-invariant RIR, we compare the proposed approach to the Lebart's method (Lebart et al., 2001), the correlation-based method in Erkelens and Heusdens (2010), the improved model-based method in Bao and Zhu (2013) and the true (perfect) LRSV estimator. The Lebart's method is actually a special case of the scheme in (10) with $\kappa(k) = 1$. The correlation-based method, as expressed by Eq. (26) in Erkelens and Heusdens (2010), is based on obtaining $\hat{X}_R(k,l)$ by (20) and then smoothing it to estimate the LRSV. Yet, due to the unavailability of long-term expectations in (22), this method uses a recursive smoothing scheme to find the prediction coefficients $c_q(k)$. The improved model-based method in Bao and Zhu (2013) uses more than one term of the past spectral variances of the reverberant speech in order to obtain a smoother shape parameter and is in fact an extension of the model-based method in Erkelens and Heusdens (2010). The latter, as expressed by Eq. (51) in Erkelens and Heusdens (2010), exploits past estimates of the LRSV averaged over frequency bins to obtain the shape parameter $\kappa(l)$. It should be noted that the correlation-based and model-based methods in Erkelens and Heusdens (2010) are developed respectively for time-invariant and time-variant RIRs. Finally, the true LRSV, which is used as a reference for comparison, is obtained by temporal smoothing of the late reverberant magnitude spectrum. The latter can in turn be calculated by convolving the anechoic speech with the late component of the RIR, i.e. that excluding the first 60 ms.

[5] Only the RIR at the first channel is considered as the observation herein.

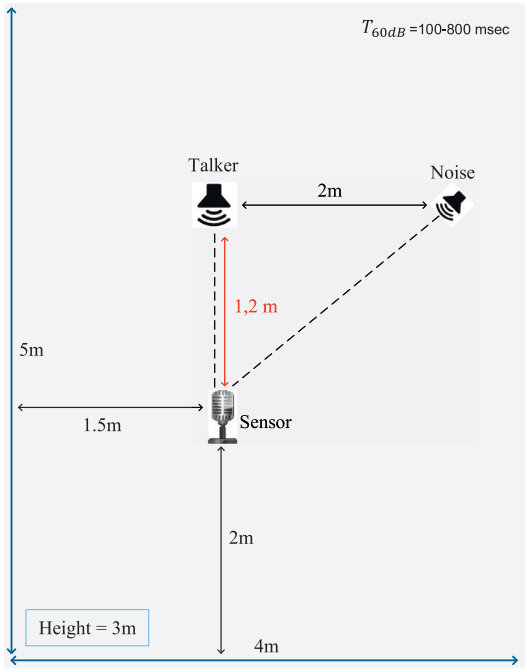

Fig. 3. A two-dimensional schematic of the geometric setup used to synthesize the time-invariant RIR by the ISM method.

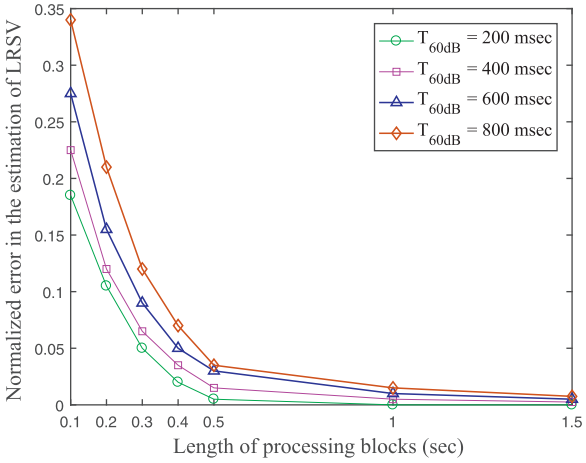

Fig. 4. Normalized error in the estimation of the LRSV w.r.t. to the case of using the entire speech utterance, versus the processing block length for different reverberation times.

To evaluate the efficiency of the proposed method w.r.t. the length of processing blocks, we calculate a measure of the error in the estimation of LRSV versus the block length for different reverberation times, as shown in Fig 4. For this figure, we consider a 3 s speech segment using processing block lengths of 0.1 to 1.5 s to estimate the LRSV and calculated the following normalized error

$$e(\Delta) = E_l\left\{\frac{||\hat{\bar{\sigma}}^2_{X_L}(k,l,\Delta) - \bar{\sigma}^2_{X_L}(k,l)||_2}{||\bar{\sigma}^2_{X_L}(k,l)||_2}\right\} \tag{30}$$

where $\hat{\bar{\sigma}}^2_{X_L}(k,l,\Delta)$ and $\bar{\sigma}^2_{X_L}(k,l)$ respectively denote the estimated LRSV using a block length of Δ and that using the entire speech utterance, $||.||_2$ is the ℓ_2-norm over frequency bins and $E_l\{.\}$ is the expected value over frames. As observed in Fig. 4, for a processing block length of 0.5 s, the relative error in the estimation of LRSV

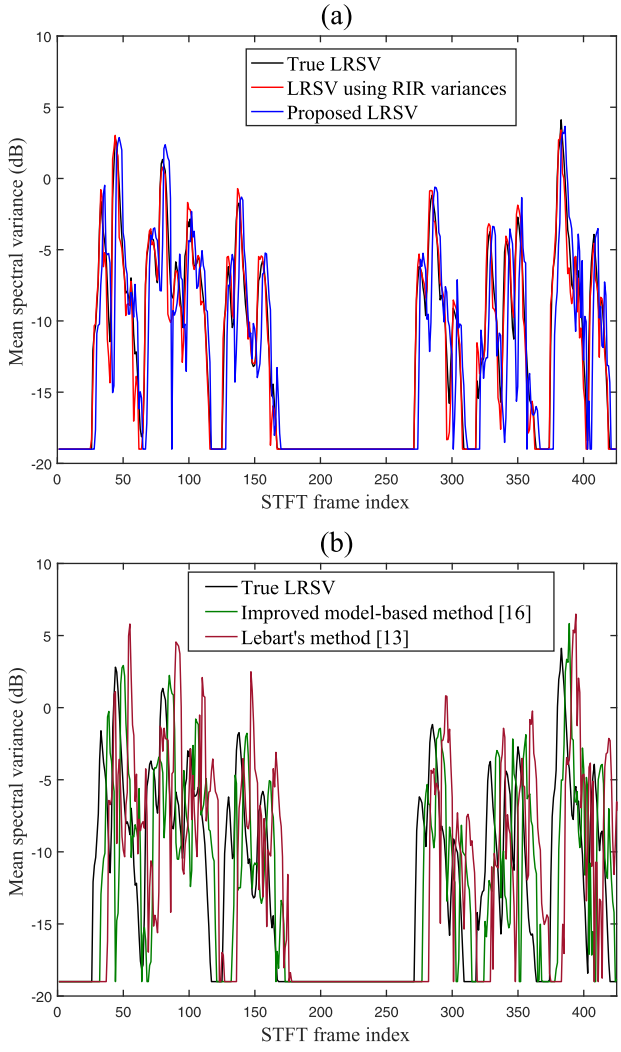

Fig. 5. Mean spectral variances using the recorded RIR from REVERB Challenge (Kinoshita et al., 2013) for: (a) the true LRSV, the LRSV estimated using RIR variances and the proposed LRSV (b) the true LRSV, the LRSV estimated by the improved model-based method (Bao and Zhu, 2013) and the one estimated by Lebart's method (Lebart et al., 2001).

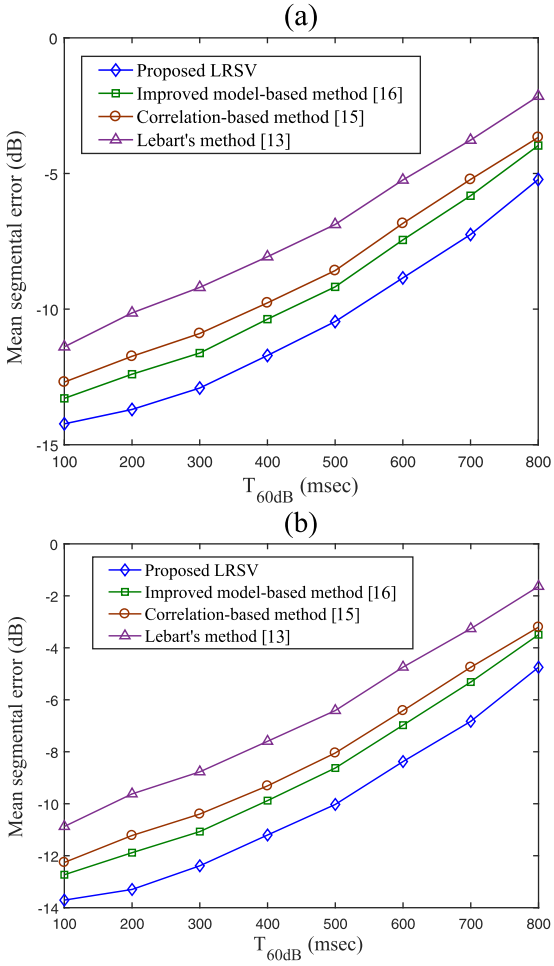

Fig. 6. Mean segmental error for different LRSV estimators using the synthesized RIRs by the ISM method Lehmann with a source-to-microphone distances of (a): 1 m (b): 2 m.

is much smaller than that for shorter blocks of around 0.1 to 0.2 s, and is almost close to that for longer blocks of 1 or even 1.5 s. In fact, even though choosing a longer processing block reduces the error defined in (30), due to the processing delay imposed by the block length, a trade-off has to be considered in the choice of the block length.

Next, to determine how close the estimated LRSVs are w.r.t. the true LRSV, we investigate the mean spectral variances, which are obtained by averaging the LRSVs over all frequency bins, as in Habets et al. (2009). The results are illustrated in Fig. 5 for an speech utterance of 425 frames. All mean spectral variance values are lower thresholded for better illustration. In order to examine how fast the methods can track abrupt changes in LRSV, we consider a short period of deactivation for the anechoic speech around the middle of the utterance. In Fig. 5 (a), the mean spectral variance of the proposed LRSV compared to that of the true LRSV and the LRSV obtained by using the knowledge of RIR variances are shown. The latter, which is used as another reference method for comparison, is obtained by using the available RIR, i.e. $h(n)$, as the following. First, we calculate the DRR as

$$\mathrm{DRR}(k) = \frac{\sum_{m=0}^{N_E-1}[h(m)]^2}{\sum_{m=N_E}^{L_h-1}[h(m)]^2}\ , \quad \forall k \tag{31}$$

and then we use the calculated DRR in (11) to obtain the shape parameter κ. The latter is next used in the smoothing scheme in (10), as in Habets et al. (2009). It is observed that the proposed LRSV is able to closely track the true LRSV and the one by using RIR variances, even in the duration of the abrupt drops/rises. As seen in Fig. 5 (b), the Lebart's (Lebart et al., 2001) and the improved model-based (Bao and Zhu, 2013) methods, still follow the LRSV but with larger errors and more delays w.r.t. the true LRSV.

Next, to evaluate the error in the proposed and considered LRSV estimation methods w.r.t. the true LRSV estimate, we calculate the mean segmental error for different reverberation times, as shown in Fig. 6. The mean segmental error can be computed by Habets et al. (2009)

$$\mathrm{Err}_{\mathrm{seg}} = E_l\left\{\frac{E_k\left\{|\hat{\sigma}^2_{X_L}(k,l) - \sigma^2_{X_L}(k,l)|^2\right\}}{E_k\left\{|\sigma^2_{X_L}(k,l)|^2\right\}}\right\} \tag{32}$$

where $\hat{\sigma}^2_{X_L}(k,l)$ and $\sigma^2_{X_L}(k,l)$ are respectively the estimated and true LRSVs, and $E_l\{.\}$ and $E_k\{.\}$ respectively denote the expectation over frames and frequency bins. As seen in Fig. 6, for both source-to-microphone distances of 1 m and 2 m, the proposed LRSV esti-

Table 1
Performance measures using the recorded RIR from REVERB Challenge.

Method	PESQ	CD	FWSNR	SRMR
Unprocessed	1.87	4.97	3.64	4.04
True LRSV	2.25	4.40	6.70	6.74
Proposed method	2.13	4.61	5.89	5.91
Method in Bao and Zhu (2013)	2.03	4.82	5.26	5.58
Method in Erkelens and Heusdens (2010)	1.97	4.88	5.10	5.52
Lebart's method (Lebart et al., 2001)	1.88	5.03	4.65	5.11

mator attains smaller errors in the entire range of T_{60dB}, as compared to the other methods.

In order to evaluate the reverberation suppression achieved by exploiting the proposed LRSV estimation approach in a gain function-based SE method, we employ the popular Bayesian log-spectral amplitude (LSA) gain function in Ephraim and Malah (1985). This scheme tends to perform late reverberation suppression using the true and estimated LRSVs. The *a priori* signal-to-reverberation ratio (SRR) required by the gain function is estimated by the DD approach (Ephraim and Malah, 1984), and to obtain the best subjective performance, the LSA gain function was lower bounded to −10 dB. In Table 1, the four aforementioned performance scores have been respectively shown for the unprocessed speech and the enhanced one by using the true LRSV, proposed LRSV, improved model-based method (Bao and Zhu, 2013), correlation-based method (Erkelens and Heusdens, 2010) and Lebart's method (Lebart et al., 2001). The results are obtained by using the recorded RIR from the REVERB Challenge dataset (Kinoshita et al., 2013). Furthermore, the same performance scores have been reported in Table 2 for the case of synthetic RIRs using the ISM method with T_{60dB} changing from 200 ms to 800 ms and the source-to-microphone distance of 1 m. It is seen that the proposed method is able to achieve the closest performance to the true LRSV as compared to the others. While the improved model-based method performs slightly better than the correlation-based method, the Lebart's method has the lowest scores. Furthermore, it can be inferred that as T_{60dB} is increased, the performance of all LRSV estimation methods degrades w.r.t. that of the true LRSV, indicating that the estimation of LRSV is a more challenging problem for highly reverberant environments. This is consistent with the results obtained for the mean segmental error in Fig. 6. Table 3 shows the same trend for a source-to-microphone distance of 2 m, resulting in slightly degraded performance as compared to Table 2. It is found that the relative performance of the considered methods in terms of the four investigated scores is consistent.

4.3. Performance in time-varying RIRs

In this part, we evaluate the relative performance of the proposed LRSV estimation method in a scenario where the RIR is time-variant. In Fig. 7, an illustration of this scenario where the ISM method is used to generate the corresponding impulse responses is shown. As seen, a talker is moving from the initial point at t=0 to the ending position at t=10 s along a straight line, resulting in a variable impulse response for the source-to-microphone channel. Here, we estimate the continuous trajectory by 20 discrete points and obtain the corresponding RIR for each point by using the ISM method. Next, a 10 s anechoic speech utterance is segmented into 20 utterances and the resulting utterances are filtered by the generated RIRs at the discrete points. The entire reverberant speech sample is generated next by combining the 20 individual segments. In this way, the continuous trajectory is well approximated by the 20 discrete points.

In Fig. 8, the mean spectral variances are shown for the true LRSV, the one obtained by the knowledge of RIR variances, the estimated LRSV by the proposed and other methods. It is observed that, whereas the proposed method is able to follow the true LRSV with visibly good precision, the other indicated methods track the changes in the true LRSV with a higher error which becomes larger in the location of abrupt decays and rises. Yet, the proposed LRSV estimator proves to be more robust against the abrupt changes in the LRSV values due to its adaptation with the changing RIR.

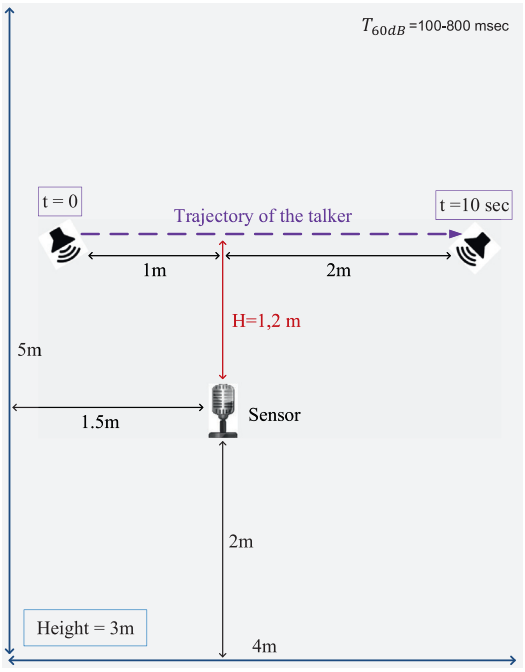

Fig. 7. A two-dimensional schematic of the geometric setup used to synthesize the time-variant RIR (moving talker) by the ISM method.

Next, we evaluate the mean segmental error in (32) in case of the time-varying RIR for the proposed method along with the improved model-based method (Bao and Zhu, 2013), model-based method (Erkelens and Heusdens, 2010) and Lebart's method (Lebart et al., 2001). As observed in Fig. 9, the same trend as that for the time-invariant RIR applies for the proposed method achieving the closest result to the true LRSV, whereas the model-based and improved model-based methods provide almost the same results particularly at higher reverberation times.

We further evaluate the reverberation suppression performance of the proposed and other methods in terms of the four aforementioned objective performance scores. In this respect, we consider the time-variant RIR scenario in Fig. 7 and compare our LRSV estimation method with the other methods. The results have been reported in Table 4 for the vertical distance H in Fig. 7 to be 1 m. Based on these results, it can be inferred that in general, the performance scores of all methods falls below those in case of time-invariant RIR. Consistent with all the performance scores, it is observed that the proposed method achieves considerably closer scores to those obtained by the true LRSV, even in higher reverberation times where the performance scores of the other methods are further degraded. This shows the advantage of the proposed method especially for changing environments. Also, it is seen that while the model-based and improved model-based methods result in almost same scores, the performance of the Lebart's method, i.e. that with a constant shape parameter, is deteriorated further than that in the case of time-invariant RIR. This shows the importance of adapting the shape parameter to the changing RIR in the estimation of LRSV.

Table 2
Performance measures using the ISM method for a source-to-microphone distance of 1 m.

PESQ				
T_{60dB} (ms)	200	400	600	800
Unprocessed	2.31	2.14	1.92	1.78
True LRSV	2.83	2.61	2.37	2.16
Proposed method	2.75	2.48	2.21	1.97
Improved model-based (Bao and Zhu, 2013)	2.71	2.43	2.14	1.90
Correlation-based (Erkelens and Heusdens, 2010)	2.70	2.41	2.12	1.88
Lebart's method (Lebart et al., 2001)	2.63	2.32	1.99	1.81
CD				
T_{60dB} (ms)	200	400	600	800
Unprocessed	3.72	4.06	4.65	5.48
True LRSV	3.03	3.39	4.11	5.06
Proposed method	3.12	3.51	4.26	5.24
Improved model-based (Bao and Zhu, 2013)	3.18	3.59	4.34	5.33
Correlation-based (Erkelens and Heusdens, 2010)	3.20	3.63	4.37	5.36
Lebart's method (Lebart et al., 2001)	3.26	3.73	4.48	5.44
FWSNR (dB)				
T_{60dB} (ms)	200	400	600	800
Unprocessed	6.03	5.12	4.16	3.04
True LRSV	9.21	8.03	6.97	5.90
Proposed method	8.69	7.32	6.27	4.95
Improved model-based (Bao and Zhu, 2013)	8.38	6.99	6.02	4.61
Correlation-based (Erkelens and Heusdens, 2010)	8.35	6.90	5.87	4.48
Lebart's method (Lebart et al., 2001)	8.02	6.64	5.49	4.16
SRMR (dB)				
T_{60dB} (ms)	200	400	600	800
Unprocessed	6.56	5.58	4.50	3.47
True LRSV	9.63	8.49	7.32	6.25
Proposed method	8.97	7.68	6.54	5.22
Improved model-based (Bao and Zhu, 2013)	8.60	7.28	6.27	4.95
Correlation-based (Erkelens and Heusdens, 2010)	8.57	7.14	6.00	4.84
Lebart's method (Lebart et al., 2001)	8.35	6.91	5.68	4.45

Table 3
Performance measures using the ISM method for a source-to-microphone distance of 2 m.

PESQ				
T_{60dB} (ms)	200	400	600	800
Unprocessed	2.28	2.12	1.87	1.75
True LRSV	2.81	2.59	2.33	2.15
Proposed method	2.72	2.46	2.20	1.94
Improved model-based (Bao and Zhu, 2013)	2.68	2.39	2.10	1.88
Correlation-based (Erkelens and Heusdens, 2010)	2.66	2.38	2.09	1.86
Lebart's method (Lebart et al., 2001)	2.60	2.29	1.96	1.78
CD				
T_{60dB} (ms)	200	400	600	800
Unprocessed	3.76	4.08	4.71	5.57
True LRSV	3.08	3.45	4.20	5.15
Proposed method	3.16	3.56	4.31	5.23
Improved model-based (Bao and Zhu, 2013)	3.21	3.63	4.40	5.39
Correlation-based (Erkelens and Heusdens, 2010)	3.24	3.67	4.46	5.42
Lebart's method (Lebart et al., 2001)	3.30	3.78	4.57	5.51
FWSNR (dB)				
T_{60dB} (ms)	200	400	600	800
Unprocessed	5.92	4.90	4.00	2.84
True LRSV	9.04	7.88	6.79	5.71
Proposed method	8.57	7.18	6.09	4.80
Improved model-based (Bao and Zhu, 2013)	8.24	6.83	5.87	4.45
Correlation-based (Erkelens and Heusdens, 2010)	8.33	7.03	5.97	4.71
Lebart's method (Lebart et al., 2001)	8.18	6.69	5.45	4.31
SRMR (dB)				
T_{60dB} (ms)	200	400	600	800
Unprocessed	6.41	5.35	4.29	3.30
True LRSV	9.52	8.30	7.19	6.08
Proposed method	8.81	7.49	6.42	5.10
Improved model-based (Bao and Zhu, 2013)	8.47	7.15	6.13	4.80
Correlation-based (Erkelens and Heusdens, 2010)	8.33	7.03	5.97	4.71
Lebart's method (Lebart et al., 2001)	8.03	6.59	5.40	4.22

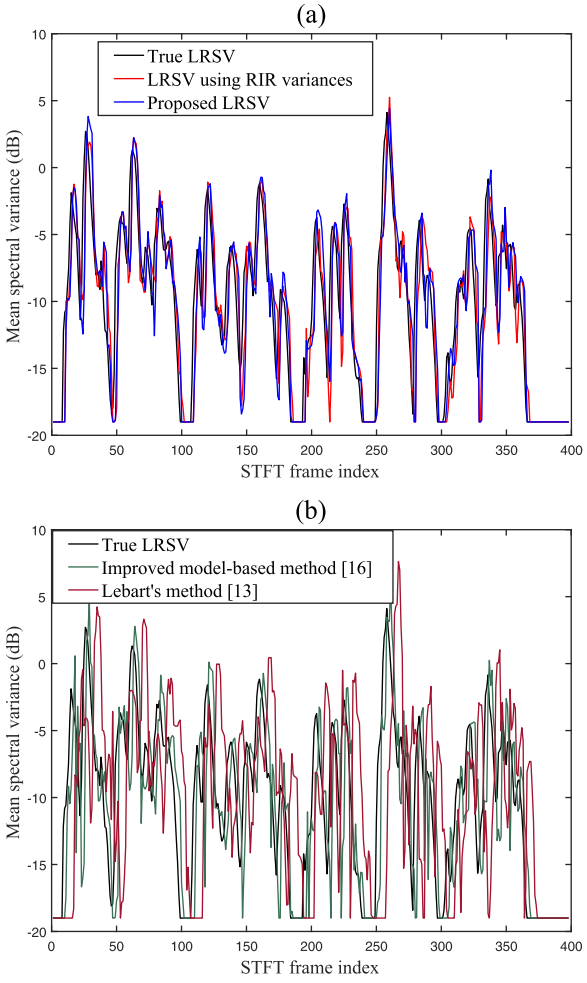

Fig. 8. Mean spectral variances for: (a) the true LRSV, the LRSV estimated using RIR variances and the proposed LRSV (b) the true LRSV, the LRSV estimated by the improved model-based method (Bao and Zhu, 2013) and the one estimated by the Lebart's method (Lebart et al., 2001).

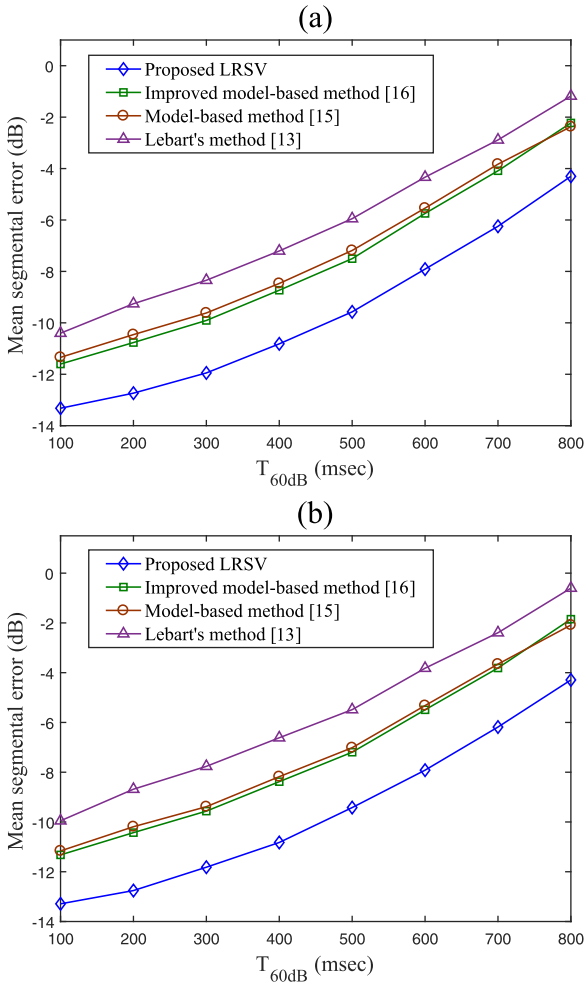

Fig. 9. Mean segmental error for different LRSV estimators using the configuration in Fig. 7 with H as (a): 1 m (b): 2 m.

In order to demonstrate the advantage of the proposed LRSV estimation method against the previous methods in real world time-varying environments, we use the RevDyn speech database available at Schwarz (2017). In this experimentation, the recordings were performed in a room with dimensions of 6 m × 5.9 m × 2.3 m and a T_{60dB} of 750 ms. The experiments involve speaking in different locations in the room and walking naturally between them. Also, there are other experiments where only slight movements such as head turning, sitting down and standing up are considered. The speaker-to-microphone distance varies between 2 m and 3.8 m. To take into account the effect of background noise, we also add babble noise to the recorded reverberant signals at different reverberant SNRs in the range of [5, 20] dB. Since the inaccuracy in the estimation of the spectral variance, LRSV, can also appear as distortion in the enhanced speech, we focus this time on the resulting distortion introduced by using each of the LRSV estimators. We here employ two frequently used measures of distortion, namely, the log-spectral distance (LSD) (Habets, 2007) and the Mel-frequency cepstral coefficients (MFCC) (Zheng et al., 2001) distance. Spectral domain measures, e.g. the LSD, are often less influenced by time misalignments between the clean and enhanced speech. We therefore use the LSD as one of the oldest distortion measures exploited for speech enhancement, which can be formed by the ℓ_p-norm of the difference between the log-STFT of the anechoic and reverberant/dereverberated signals (Habets, 2007). As well, we use the MSE between the MFCC of the anechoic and enhanced signals, to have a more complete performance assessment in the case of noisy reverberant environments, since the MFCC coefficients are rather sensitive to background noise (Zheng et al., 2001; Bao and Zhu, 2013). The MFCC distance, in addition to the audible quality of speech, is related to automatic speech recognition (ASR) performance and has been used as one of the main features for ASR systems. The corresponding results are shown in Fig. 10 versus different noise levels. The presented values are averages for three different scenarios from Schwarz (2017). As observed, the smaller LSD and MFCC distances for the proposed LRSV estimator is indicative of a lower distortion implied by using our approach, as compared to the rest of the methods. This demonstrates the advantage of the proposed method in changing environments where the RIR is time-variant. It has to be noted that, for SNR values of 5 dB and lower, the problem becomes more of a joint noise and reverberation suppression, and therefore, the performance of all employed methods tends to degrade.

The main contribution of the proposed LRSV estimation approach has two aspects: the selection of the shape parameter $\kappa(k, l)$ by (13) and the MA-based method to estimate the reverberant-only component $X_R(k, l)$ by (21). The latter method, in fact, plays an important role in the entire LRSV estimation approach by eliminating the direct-path component from the observed reverberant speech. In order to show the pure advantage with the proposed scheme for the estimation of $X_R(k, l)$, we employ the WPE-

Table 4
Performance measures for the time-variant RIR in Fig. 7 with H = 1 m.

PESQ				
T_{60dB} (ms)	200	400	600	800
Unprocessed	2.28	2.13	1.92	1.77
True LRSV	2.76	2.58	2.29	2.10
Proposed method	2.71	2.40	2.16	1.93
Improved model-based (Bao and Zhu, 2013)	2.66	2.35	2.10	1.84
Model-based (Erkelens and Heusdens, 2010)	2.66	2.36	2.09	1.83
Lebart's method (Lebart et al., 2001)	2.56	2.27	1.93	1.76
CD				
T_{60dB} (ms)	200	400	600	800
Unprocessed	3.80	4.09	4.65	5.49
True LRSV	3.16	3.54	4.26	5.28
Proposed method	3.20	3.62	4.37	5.39
Improved model-based (Bao and Zhu, 2013)	3.25	3.71	4.50	5.44
Model-based (Erkelens and Heusdens, 2010)	3.26	3.71	4.52	5.45
Lebart's method (Lebart et al., 2001)	3.31	3.77	4.61	5.52
FWSNR (dB)				
T_{60dB} (ms)	200	400	600	800
Unprocessed	5.90	5.07	4.15	3.02
True LRSV	8.97	7.82	6.69	5.63
Proposed method	8.30	7.21	6.02	4.60
Improved model-based (Bao and Zhu, 2013)	8.21	6.70	5.52	4.08
Model-based (Erkelens and Heusdens, 2010)	8.20	6.68	5.52	4.06
Lebart's method (Lebart et al., 2001)	7.92	6.43	5.18	3.81
SRMR (dB)				
T_{60dB} (ms)	200	400	600	800
Unprocessed	6.48	5.55	4.51	3.44
True LRSV	9.46	8.21	6.97	5.90
Proposed method	8.64	7.32	6.22	4.91
Improved model-based (Bao and Zhu, 2013)	8.32	6.96	5.92	4.60
Model-based (Erkelens and Heusdens, 2010)	8.30	6.92	5.84	4.57
Lebart's method (Lebart et al., 2001)	8.03	6.59	5.40	4.22

based scheme in (13) to obtain the shape parameter in the recursive smoothing step in different LRSV estimation methods from the literature. The corresponding PESQ measure for the resulting combination of the LRSV estimation methods with the scheme in (13) along with that for the proposed approach have been shown in Fig. 11 for the same scenario as Table 4. Apart from the improvement of the underlying methods as compared to Table 4, it can be observed that the proposed approach still outperforms the rest of the LRSV estimation methods, which is due to the benefit from the reverberant-only estimation scheme provided by (13).

As aforementioned, we used the LSA gain function in Ephraim and Malah (1985) as the underlying SE method to suppress the late reverberation for our experiments. In fact, we experimented that the most efficient gain function-based SE method for the suppression of late reverberation is the log-MMSE method, i.e. the LSA gain function, in Ephraim and Malah (1985) and that the other more recent similar methods did not provide further performance advantage[6]. Nevertheless, in order to show the applicability of the proposed LRSV estimation to different gain function-based SE techniques, we here present an experiment with a few of the other such techniques, namely, the traditional Wiener filtering (Loizou, 2013), a version of the spectral subtractive method (Loizou, 2013), and the MAP amplitude estimation with a super-Gaussian speech prior (Lotter and Vary, 2005). The resulting PESQ scores for the same scenario as Table 4 are presented in Fig. 12 with the proposed LRSV estimator used in different gain functions to suppress late reverberation. It can be inferred that, while the LSA gain function performs best, the rest of the methods score almost closely to this gain function when using the proposed LRSV approach.

[6] This is in contrast with the gain function-based noise reduction where there are plenty of gain functions and their modified versions, being able to provide further enhancement. For a complete literature review on this, the reader is referred to Parchami et al. (2016).

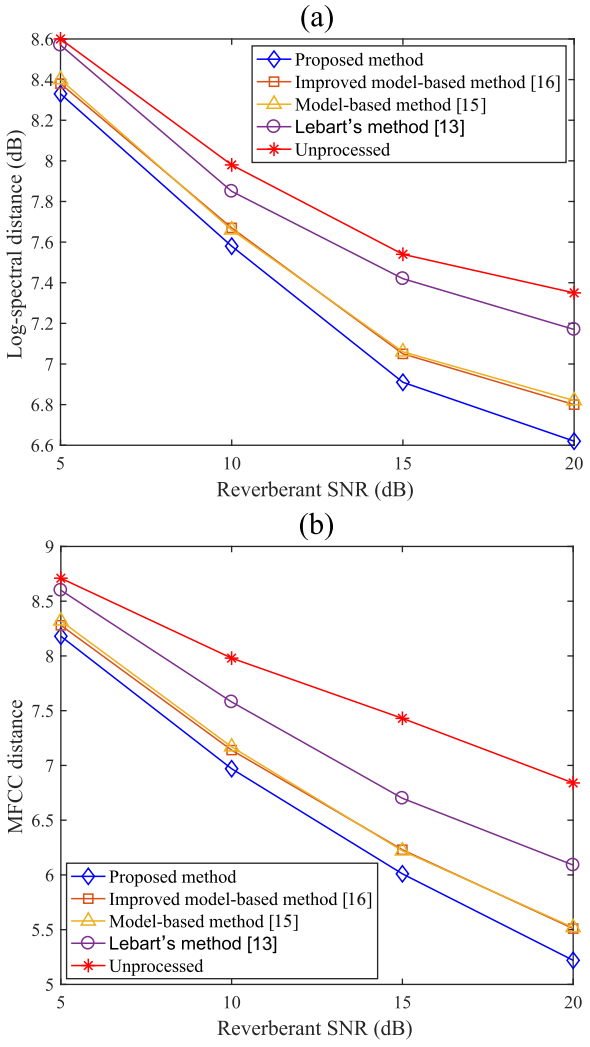

Fig. 10. (a): Log-spectral distance (lower is better) and (b): MFCC distance (lower is better), versus the reverberant SNR for the time-varying acoustic scenario in Schwarz (2017) using different methods.

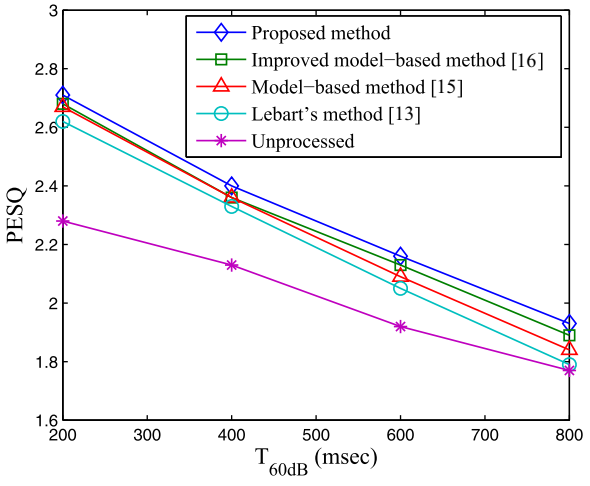

Fig. 11. Performance comparison of different LRSV estimation methods using the proposed WPE-based shape parameter in (13) in their recursive smoothing scheme.

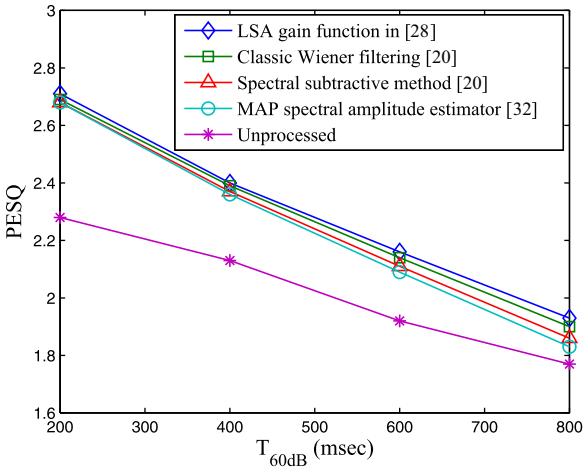

Fig. 12. Performance comparison of different gain function-based SE methods with the proposed LRSV estimation approach.

5. Conclusion

We proposed a model-based estimator for the spectral variance of the late reverberant speech using a modification of the WPE dereverberation method. The suggested approach employs the WPE method in an incremental processing manner with a short delay, where preliminary estimates of the reverberant and direct-path components of speech are extracted from each processing block. These estimates are further exploited in a model-based smoothing scheme to estimate the LRSV. We evaluated the performance of the proposed LRSV estimation method in terms of different performance measures recommended by the REVERB Challenge in both time-invariant and time-variant acoustic environments. According to the experiments, the proposed LRSV estimator outperforms the previous major methods considerably and scores the closest results to the theoretically true LRSV estimator. Particularly, in case of changing RIRs where other methods fail to precisely follow the true LRSV estimator, our estimator is able to track the true LRSV values closely. The proposed approach is performed in a blind way and does not require any prior information about the speech or acoustic parameters. Future work in this direction involves taking into account the inherent correlation of the early and late reverberant components of speech, reducing the processing block length and making the proposed approach robust against fast changes in the RIR.

Acknowledgment

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) Grant No. N01163 of Canada.

References

Recommendation P.862, 2001. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecsITU-T.
Garofolo et al., J.S., 1993. TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Philadelphia: Linguistic Data Consortium.
Bao, X., Zhu, J., 2013. An improved method for late-reverberant suppression based on statistical model. Speech Commun. 55 (9), 932–940.
Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoustics Speech Signal Process. 32 (6), 1109–1121.
Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoustics Speech Signal Process. 33 (2), 443–445.
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.
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.
Furuya, K., Kataoka, A., 2007. Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction. IEEE Trans. Audio Speech Language Process. 15 (5), 1579–1591.
Gannot, S., Burshtein, D., Weinstein, E., 2001. Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans. on Signal Process. 49 (8), 1614–1626.
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., Gannot, S., Cohen, I., 2009. Late reverberant spectral variance estimation based on a statistical model. IEEE Signal Process. Lett. 16 (9), 770–773.
Hu, Y., Loizou, P.C., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Language Process. 16 (1), 229–238.
Huang, Y., Benesty, J., Chen, J., 2005. A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment. IEEE Trans. Speech Audio Process. 13 (5), 882–895.
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: 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 1–4.
Kumar, K., Stern, R.M., 2010. Maximum-likelihood-based cepstral inverse filtering for blind speech dereverberation. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4282–4285.
Lebart, K., Boucher, J.M., Denbigh, P.N., 2001. A new method based on spectral subtraction for speech dereverberation. Acta Acoust. 87, 359–366.
Lehmann, E. A.,. Image-source method: matlab code implementationAvailable at http://www.eric-lehmann.com/.
Loizou, P.C., 2013. Speech Enhancement: Theory and Practice, Second Edition. CRC Press.
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).
Lotter, T., Vary, P., 2005. Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model. EURASIP J. Appl. Signal Process. 2005, 1110–1126.
Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., Juang, B.H., 2008. 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), 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.A., Gaubitch, N.D. (Eds.), 2010, Speech Dereverberation. Springer-Verlag, London.
Parchami, M., Zhu, W.P., Champagne, B., Plourde, E., 2016. Recent developments in speech enhancement in the short-time fourier transform domain. IEEE Circuits Syst. Mag. 16 (3), 45–77.
Polack, J.D., 1988. La transmission de l' energie sonore dans les salles. Université du Maine, La Mans, France Ph.D. thesis.
Schwarz, B., 2017. RevDyn Speech DatabaseSpeech and Acoustic Lab of the Faculty of Engineering at Bar Ilan University, Last accessed on March, Available at http://www.eng.biu.ac.il/gannot/speech-enhancement/.
Warsitz, E., Haeb-Umbach, R., 2007. Blind acoustic beamforming based on generalized eigenvalue decomposition. IEEE Trans. Audio Speech Language Process. 15 (5), 1529–1539.
Yoshioka, T., 2010. Speech Enhancement in Reverberant Environments. Kyoto University, Japan Ph.D. thesis.
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.
Zheng, F., Zhang, G., Song, Z., 2001. Comparison of different implementations of MFCC. J. Comput. Sci. Technol. 16 (6), 582–589.