Reverberation Time Estimation based on a Model for the Power Spectral Density of Reverberant Speech

Neda Faraji
Engineering and Technology Department
Imam Khomeini International University
Qazvin, Iran
Email: nfaraji@aut.ac.ir

Seyed Mohammad Ahadi
Amirkabir University of Technology
Tehran, Iran
Email: sma@aut.ac.ir

Hamid Sheikhzadeh
Amirkabir University of Technology
Tehran, Iran
Email: hsheikh@aut.ac.ir

***Abstract*—In this paper, it is shown that the Power Spectral Density (PSD) of late reverberant speech can be described by a first-order Infinite Impulse Response (IIR) model with the pole related to the reverberation time. Utilizing this first-order IIR model, an online method for reverberation time estimation (RTE) from a recorded reverberant signal is proposed. The proposed method takes advantage of processing in subband domain in order to reliably estimate the reverberation time in noisy environments. Comparing with a well-known maximum likelihood approach for RTE, the superior performance of the new approach for fast tracking of RT with higher accuracy is demonstrated.**

I. Introduction

The energy of a reverberant sound field in a room decays exponentially with a specific rate related to a parameter that is called Reverberation Time (RT_{60}). It is defined as the time interval in which the energy of the reverberant sound field decays 60 dB below its initial energy level after switching-off the excitation source [1]. Knowledge about RT is employed significantly in speech dereverberation techniques [2] and is also of interest for acousticians in architectural design of auditoriums and large chambers.

Some methods determine the RT_{60} parameter of an enclosed environment based on explicit equations related to the room geometry and absorptive properties of the objects in it [1]. In addition, RT_{60} can be estimated from the slope of a measured Room Impulse Response (RIR) by scattering either a burst of noise or brief pulse into the test enclosure [3], [4]. The measurement-based methods require careful experiments and suitable excitation signals. Besides, these offline methods are less practical as we encounter time-varying RT_{60}. Generally, entirely blind methods are preferred that work without having any prior knowledge of the received signal, room geometry and the absorptive characteristics of the objects. A well-known method in this category is [5] which estimates the reverberation time by a Maximum Likelihood (ML) approach. This ML-based approach [5] was extended in [6] for reverberation time estimation from noise-corrupted reverberant speech signal.

In this paper, we present a novel online reverberation time estimation method that can be employed in intelligibility improvement algorithms of Public Address (PA) systems where the original clean speech is available. In [7], a noise PSD estimation algorithm has been proposed for reverberant enclosures assuming that the reverberation time is known. Utilizing our proposed RT estimation method here, the noise PSD estimation approach [7] can be generalized to the case of unknown reverberation time.

To perform RTE, we utilize a new formula for the PSD of late reverberant speech [7]. In contrast to the approach [6] that use a time-domain model for the reverberant speech within the pause intervals, the derived PSD-domain model holds for any arbitrary segment of the reverberant speech. Through fitting the theoretical reverberant PSD to the observed PSD, RT_{60} can be inferred segment-by-segment. Moreover, the algorithms extracting RT_{60} only in the free decay parts of the reverberant speech require a long recorded signal, to provide a large number of free-decay parts. In this paper, by exploiting a segment-by-segment strategy for continuous RTE, the need for long data record is surmounted and fast tracking of RT_{60} would be possible. Furthermore, by performing RT estimation in the subband domain it is possible to remove the noise-dominated PSD bins for reliable estimation of RT_{60} in noisy environments. We demonstrate that our proposed method outperforms Löllmann's approach [6] in both accuracy and speed of tracking RT_{60}.

This paper is organized as follows. The next section is devoted to the basic assumptions and observation model of a recorded reverberant speech. We present the proposed method in Sec. III. Then, the simulation setup and experimental results are explained in Sec. IV. Finally, we conclude the paper in Sec. V.

II. Observation Model

Assume a clean message is played through a loudspeaker (source) in a reverberant environment. Let s, x and d denote the played clean speech, the late reverberant version of s, and the noise signal respectively. The signal y recorded by the

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

microphone at a specified distance from the source is defined by the following equation

$$y(n) = \alpha s(n - n_0) + x(n) + d(n), \quad (1)$$

where n is the discrete-time index, and α and n_0 are the attenuation factor and sound propagation delay between the source and receiver, respectively. Assuming a specified distance between the source and receiver, n_0 can be removed to simplify the model (1). With this zero-delay assumption, the late reverberant clean speech signal x can be rewritten as

$$x(n) = h(n-1) * s(n), \quad (2)$$

in which $*$ is the convolution operator and $h(n)$ represents the Room Impulse Response (RIR). According to the Polack's model, a RIR is generated as one realization of the following stochastic process [7]

$$h(n) = b(n)e^{-\eta n}\, n \geq 0, \quad (3)$$

where $b(n)$ is a zero-mean normally distributed stochastic process with variance ν^2, which defines the fine structure of the RIR modulated with an exponential function with the decay rate of η. The decay rate is defined by $\eta = \frac{3\ln(10)}{RT_{60} f_s}$ in which RT_{60} and f_s are the reverberation time and sampling frequency, respectively. Assuming frame-by-frame processing, for the i^{th} frame and k^{th} Discrete Fourier Transform (DFT) bin, we can rewrite (1) in the frequency domain as $(n_0 = 0)$

$$Y(i,k) = \alpha S(i,k) + X(i,k) + D(i,k). \quad (4)$$

Assuming α and s are known, we can refer to the observed PSD by $Z(i,k)$ which is defined as

$$Z(i,k) = Y(i,k) - \alpha S(i,k) = X(i,k) + D(i,k). \quad (5)$$

III. PROPOSED METHOD

A. Model of the reverberant PSD

Faraji et. al showed that the PSD value of late reverberant speech at i^{th} frame and k^{th} DFT bin is determined by [7]

$$\sigma_X^2(i,k) = \sum_{p=0}^{\infty} \nu_{i+\frac{N}{2}}^2 e^{-2\eta_{i+\frac{N}{2}} p} S^2(i-p-1,k), \quad (6)$$

in which $S(i-p-1,.)$ denotes the DFT of a frame which starts at $(i-p-1)^{th}$ sample index of the clean speech signal. Also $\nu_{i+\frac{N}{2}}$ and $\eta_{i+\frac{N}{2}}$ are time-varying parameters of the RIR according to the Polack's model, and N is the frame length. As the PSD (6) non-linearly depends on the decay rate parameter, in [8] Faraji et. al utilized a Non-Linear Least Squares (NLLS) method to estimate the reverberation time in noise-free reverberant enclosures. Due to the high complexity of running NLLS method in each frequency bin, in this paper we simplify the proposed model of (6) to reduce the computational burden of estimating reverberation time.

B. IIR model of the reverberant PSD

we can rewrite (6) as

$$\begin{aligned} \sigma_X^2(i,k) &= \nu_{i+\frac{N}{2}}^2 S^2(i-1,k) \\ &+ \sum_{p=1}^{\infty} \nu_{i+\frac{N}{2}}^2 e^{-2\eta_{i+\frac{N}{2}} p} S^2(i-p-1,k), \end{aligned} \quad (7)$$

By change of variable $q = p - 1$, we can simplify (7) as

$$\begin{aligned} \sigma_X^2(i,k) &= \nu_{i+\frac{N}{2}}^2 S^2(i-1,k) \\ &+ \sum_{q=0}^{\infty} \nu_{i+\frac{N}{2}}^2 e^{-2\eta_{i+\frac{N}{2}} (q+1)} S^2(i-q-2,k) . \end{aligned} \quad (8)$$

Assuming that the parameters of RIR are time-invariant, we can remove the subscript $i + \frac{N}{2}$ from $\nu_{i+\frac{N}{2}}^2$ and $\eta_{i+\frac{N}{2}}$. Finally, we can rewrite (8) as follows

$$\begin{aligned} \sigma_X^2(i,k) &= \nu^2 S^2(i-1,k) \\ &+ e^{-2\eta} \sum_{q=0}^{\infty} \nu^2 e^{-2\eta q} S^2(i-q-2,k) \\ &= e^{-2\eta}\, \sigma_X^2(i-1,k) + \nu^2 S^2(i-1,k). \end{aligned} \quad (9)$$

To be practically tractable, the summation in (6) is limited to an upper bound which corresponds to considering a finite length RIR. For a limited upper bound Q, we obtain (10) in place of (9)

$$\begin{aligned} \sigma_X^2(i,k) &= e^{-2\eta}\, \sigma_X^2(i-1,k) + \nu^2 S^2(i-1,k) \\ &- \nu^2\, e^{-2\eta(Q+1)}\, S^2(i-Q-2,k). \end{aligned} \quad (10)$$

Meanwhile, assuming that Q is large, which is a plausible assumption for RIRs, we can discard the last term in (10). To summarize, we showed that the time sequence of late reverberant PSD points in each DFT bin can be modelled by a first-order IIR with $e^{-2\eta}$ as the pole location.

C. Decay rate estimation method

Let us consider a short-time segment of late reverberant speech which starts at n^{th} discrete-time index with the length of N samples, i.e., $\{x(n), x(n+1),\ ..., x(n+N-1)\}$. This segment is divided into the frames of L samples $(L < N)$ and the frame shift of a single sample. We denote the PSD points of m^{th} frame in the segment by $\sigma_X^2(m,k)$ for $1 \leq m \leq M$. According to (9), the decay rate estimate can be obtained as

$$\hat{\eta}(j) = \underset{\eta}{min}\;\; f(\eta, j), \quad (11)$$

where $f(\eta, j)$ is the objective function to be minimized

$$f(\eta, j) = \sum_{k=1}^{N_F/2+1} \sum_{m=2}^{M} e^2(m,k), \quad (12a)$$

$$e(m,k) = \sigma_X^2(m,k) - e^{-2\eta}\sigma_X^2(m-1,k) - \nu^2 S^2(m-1,k), \quad (12b)$$

in which N_F is the number of DFT coefficients and j is the segment number. Taking the deriviate of (12a) with respect to η, $\hat{\eta}(j)$ would be

$$\hat{\eta}(j) = -0.5 \log\Bigg\{ \frac{\sum_{k=1}^{\frac{N_F}{2}+1} \sum_{m=2}^{M} \sigma_X^2(m-1,k)\left[\sigma_X^2(m,k) - \nu^2 S^2(m-1,k)\right]}{\sum_{k=1}^{\frac{N_F}{2}+1} \sum_{m=2}^{M} \sigma_X^4(m-1,k)} \Bigg\} \quad (13)$$

Although, the expression in (13) can simply determine the η parameter, in application rather than the true PSD, only an estimate of the PSD of reverberant speech, i.e., $\hat{\sigma}_X^2(i,k)$ is available. It is well known that the Least Square (LS) estimate of the coefficient in a first-order difference equation, like (9), is severely downward biased for the coefficients near to one [10]. The same situation happens for typical values of the reverberation time where the coefficient ($e^{-2\eta}$) is near to one. To surmount the challenge of the estimation noise in the observed PSD, first the primary decay rates of a number of J consecutive segments are estimated by (13). The secondary decay rate estimate of the j^{th} segment is obtained by inferrring some statistics, for example median or mean, from the histogram made by the primary estimated decay rates, i.e. $\hat{\eta}_{pri}(j), \hat{\eta}_{pri}(j-1), ..., \hat{\eta}_{pri}(j-J+1)$. Finally, the secondary estimate is smoothed using a first-order recursive averaging filter with the time constant of 0.996 to form the final decay rate estimate

$$\hat{\eta}_{fin}(j) = 0.996\hat{\eta}_{fin}(j-1) + 0.004\hat{\eta}_{sec}(j). \quad (14)$$

Although this approach can remove the spurious estimates, taking advantage of adaptive filtering permits us to more effectively reduce the severely-biased estimates, experimentally we found. Indeed, an adaptive strategy for parameter estimation leads to the estimation noise of σ_X^2 to be smoothed and the estimation procedure to have a stable behaviour. Employing a gradient based method to minimize the objective function (12a), we can update the j^{th} estimate of the η parameter as follows

$$\hat{\eta}(j) = \hat{\eta}(j-1) - 2\mu \sum_{m=2}^{m=M} \sum_{k=1}^{N_F/2+1} e(m,k)\frac{\partial e(m,k)}{\partial \eta}, \quad (15)$$

in which $\frac{\partial e(m,k)}{\partial \eta} = 2e^{-2\eta}\hat{\sigma}_X^2(m-1,k) \geq 0$ and μ is the learning rate. The following modified version of (15) is used in this paper

$$\hat{\eta}(j) = \hat{\eta}(j-1) - \frac{\mu}{\nu^2\, Var\{s(n)\}} \sum_{m=2}^{m=M} \sum_{k=1}^{N_F/2+1} e(m,k). \quad (16)$$

The first modification, which is normalizing the learning rate μ by $\nu^2\, Var\{s(n)\}$, is similar to the approach used in the Normalized Least Mean Squares (NLMS) method to have a constant learning rate μ for all power levels of an input signal. Moreover, experimentally we found that using the sign of the derivative term ($\frac{\partial e}{\partial \eta}$) instead of its value caused the adaptation algorithm (15) to be more stable. Therefore, the second modification is replacing the error derivative by its sign, which is similar to the approach used in the sign regressor version of an LMS algorithm [11] to increase stability. By using this strategy, we can also reduce the computational complexity of the adaptation equation (15). The secondary and final decay rate estimates are obtained by the same formulas as mentioned before.

D. Decay rate estimation from noisy reverberant speech

For a noise-corrupted reverberant speech, $\hat{\sigma}_X^2$ is replaced with the observed PSD $\hat{\sigma}_Z^2$ in (12b). The highly noise-corrupted PSD points for computing the objective function (12a) can lead to a large error in the η estimates. More accurate estimates can be obtained by incorporating the *a priori* Signal to Noise Ratios (SNR) for the PSD points. Then, only those PSD points having *a priori* SNRs greater than a predefined threshold (SNR_{th}) are employed for the estimation procedure. The *a priori* SNR of the $(m,k)^{th}$ PSD point is defined by $snr(m,k) = 10\log_{10}\frac{\hat{\sigma}_X^2(m,k)}{\hat{\sigma}_D^2(m,k)}$ in which $\hat{\sigma}_D^2(m,k)$ is an estimate of the noise PSD and $\hat{\sigma}_X^2(m,k)$ is defined by (6). However, the exact value of $\hat{\sigma}_X^2(m,k)$ is unknown, as it depends on the uncertain decay rate η to be estimated. Hence, using a rough estimate of $\hat{\sigma}_X^2(m,k)$ the approximated *a priori* SNR parameter, $asnr$, is defined by

$$asnr(m,k) = 10\log_{10}\frac{\sum_{p=0}^{P-1}\nu^2 S^2(m-p-1)}{\hat{\sigma}_D^2(m,k)}, \quad (17)$$

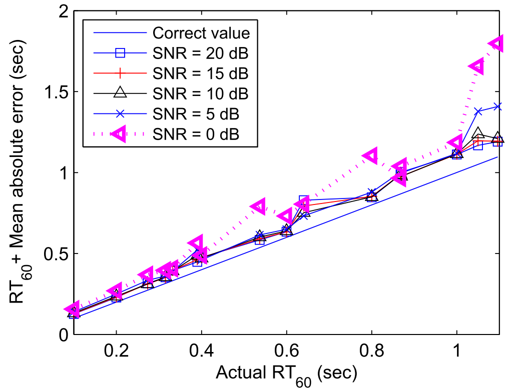
in which P is the length of RIR. The nominator in (17) is $\hat{\sigma}_X^2(m,k)$ excluding the exponential term $e^{-2\eta p}$, so the approximated *a priori* SNR, $asnr$, is always greater than the snr value. Taking into account the aforementioned idea, the objective function (12a) is modified as

$$f_{asnr}(\eta,j) = \sum_{(m,k)\in\{(m_a,k_a)|asnr(m_a,k_a)>SNR_{th}\}} e^2(m,k). \quad (18)$$

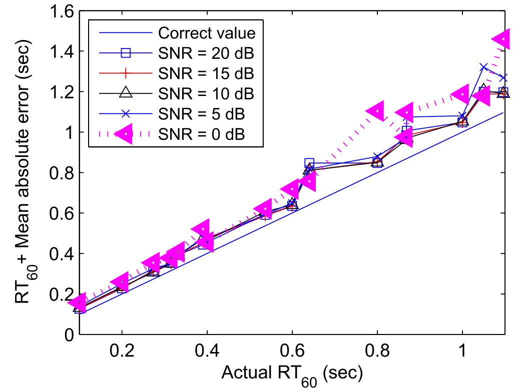
IV. SIMULATIONS AND EXPERIMENTAL RESULTS

A. Simulation setup

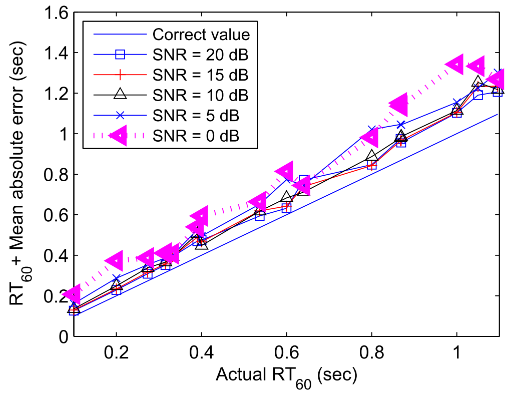
The experimental results are reported on an anechoic speech signal constituting 6 speech files from TIMIT database with the sampling frequency of 8 kHz (17-sec speech signal). To simulate the late reverberant speech signal, first we generated 6 synthetic RIRs using Polack's model with $\nu^2 = 0.1$, and $RT_{60} = (0.1, 0.2, 0.4, 0.6, 0.8, 1)$ s. In addition, we used 10 measured RIRs from Aachen Impulse Response (AIR) database [12]. The selected RIRs from AIR database constitute 5 RIRs used in the evaluation in [6] plus 5 additional ones. The reverberation times of these RIRs, measured by the Schroeder method [4], are $RT_{60} = (0.39, 0.64, 0.315, 0.87, 1.096, 0.273, 0.536, 0.33, 0.86, 1.05)$ s. In our proposed method, we assume the variance parameter in Polack's model, ν^2, to be known. Hence, first we estimate the variance ν^2 from the recorded RIRs and then apply the estimated variance into the reverberation time estimation procedure.



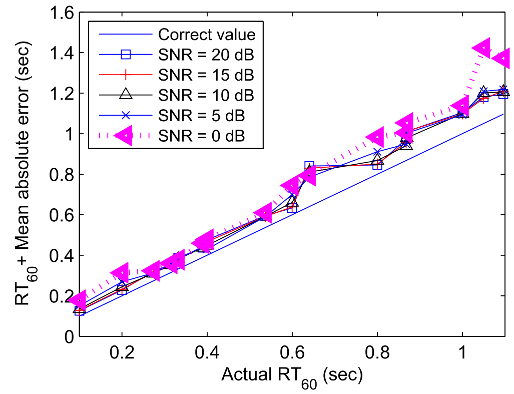
(a)



(b)



(c)



(d)

Fig. 1: RT_{60}+ Mean absolute error at different noisy conditions (a) white noise, (b) modulated-white noise, (c) babble noise, and (d) non-stationary train noise.

The obtained ν^2 estimates for the recorded RIRs are $\nu^2 = (1.3, 1.6, 2.45, 1.32, 0.00018, 1.5, 1.13, 2.51, 1.84, 0.00027) \times 10^{-4}$. The reverberant speech signals are contaminated by white noise, modulated white noise with the setup as in [13], babble noise and non-stationary train noise at a SNR level between 0dB and 20dB. Our experimental results indicated that the best performance is attained at $N = 1024$ with the segment shift of 3 ms, $J = 500$, $SNR_{th} = 20$ dB, and inferring the mean statistic from the histogram of the primary decay rates. Each segment is partitioned into the 128-sample frames weighted with a Hamming window and 127-sample overlaps. Moreover, two more true η is considered as the initial estimate and the μ parameter is set to 3×10^{-7}. The corresponding reverberation time estimate is calculated by $\hat{RT}_{60}(j) = \frac{3\ln(10)}{\hat{\eta}_{fin}(j) f_s}$. We measure the estimation performance using mean absolute error as follows

$$\text{Mean absolute error} = \frac{1}{N_{seg}} \sum_{j=1}^{N_{seg}} \left| \hat{RT}_{60}(j) - RT_{60} \right|, \quad (19)$$

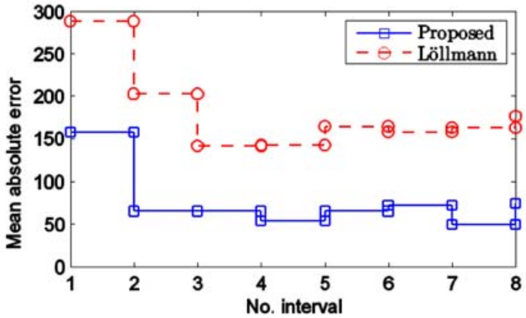
where $|.|$ represents the absolute value operator, N_{seg} denotes the total number of segments, and RT_{60} and $\hat{RT}_{60}$ represent the true and estimated reverberation times, respectively. Finally, the relative error is determined by

$$\text{Relative error} = \frac{\text{Mean absolute error}}{RT_{60}} \times 100\%. \quad (20)$$

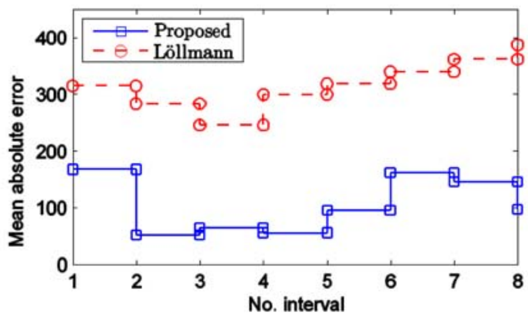
To calculate the $asnr$ to be compared with SNR_{th}, a method to provide a rough noise estimate has been employed which estimates the PSD of noise during the initial silent interval of the reverberant speech.

B. Experiment 1

In this experiment, we evaluate the performance of our proposed method at different stationary or non-stationary noise types with variable SNR levels. In each sub-figure of Fig. 1, the true RT value has been plotted as a solid line. The distance of each line with the markers from the solid line shows the Mean absolute error. As it is clear from this figure, although a rough noise estimate has been employed to calculate the $asnr$, the performance degradation is not considerable for low-SNR compared to high-SNR cases. Indeed, by performing RT estimation in the DFT domain the low-SNR PSD points could be simply removed from the RT estimation procedure. As illus-



(a)



(b)

Fig. 2: Average of Mean absolute errors (ms) for the Löllmann and proposed methods (a) 15 dB, (b) 5 dB.

trated in Fig. 1, performance degradation is not considerable for the low-SNR case compared to the high-SNR one. Our approach is similar to the recently proposed method [9] where the full-band RT of a clean reverberant speech is estimated from the subband derived RTs. However, this approach, which estimates the RT values just in the free-decay parts of the reverberant speech, is expected to fail for low-SNR conditions, as the free-decay parts of the reverberant speech are the most spurious parts. In our method, utilizing a general model for the PSD of reverberant speech, the high-SNR PSD points could be employed for RTE.

C. Experiment 2

To show that our approach is able to eliminate the need for a long record of reverberant speech to infer the RT value, we compare our method with the one proposed by Löllmann et. al. [6]. The modulated white noise is used to contaminate the reverberant speech signals at the two SNR levels of 15 dB and 5 dB. The average of mean absolute errors over 16 different reverberant speech signals at two-second intervals are shown in Fig. 2. As it is demonstrated in Fig. 2(a), our approach has the highest error in the first interval. As the time elapses, the mean absolute error reduces and remains approximately constant. More or less the same phenomenon is observed in Fig. 2(b) except in the fifth and sixth intervals during which the error increased. The intervals segmental SNR values at the global SNR level of 5 dB are 2.4, 7.3, 3.2, 9.3, -1.8, -3.9, 5.4 and 8 dB, respectively. In fact, the performance degradation has been occurred within the low-SNR intervals of the reverberant speech. In comparison, Löllmann's method has demonstrated a reasonable performance at the SNR level of 15 dB, Fig. 2(a). However, since this method estimates the RT during the free decay regions, a significant performance degradation is observed in all intervals of Fig. 2(b) which corresponds to the low-SNR case.

V. CONCLUSION

In this paper, we showed that the time sequence of late-reverberant PSD in each DFT bin (temporal envelope) follows a first-order difference equation. Using this first-order model, in which the unknown coefficient is related to the reverberation time, we were able to continuously extract reverberation time from the recorded reverberant speech. The performance of our proposed approach was evaluated in four noisy conditions. By exploiting subband processing and removing highly-contaminated PSD points, a small degradation in performance was observed for the low-SNR cases in comparison with the high-SNR ones. Moreover, our continuous estimation approach demonstrated a superior performance for the online estimation of RT only with a few seconds of the reverberant speech signal. As a suggestion for the future work, our algorithm can also be extended to the estimation of ν^2 parameter.

REFERENCES

[1] H. Kuttruff, *Room Acoustics*, Taylor & Francis, London, Fourth Edition, pp. 97–98, 127–141, 2000.

[2] S. Mosayyebpour, H. Sheikhzadeh, T. A. Gulliver, and M. Esmaeili, "Single-microphone LP residual skewness-based approach for inverse filtering of room impulse response," *IEEE Trans Audio, Speech, and Language Processing* **20**, pp. 1617–1632, 2012.

[3] *ISO-3382: Acoustics-Measurement of the Reverberation Time of Rooms with Reference to Other Acoustical Parameters*, International Organization for Standardization, Geneva, Switzerland, 1997.

[4] M. R. Schroeder, "New method of measuring reverberation time," *Journal of the Acoustical Society of America* **37**, pp. 409–412, 1965.

[5] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. OâĂŹBrien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *Journal of the Acoustical Society of America* **114**, pp. 2877–2892, 2003.

[6] H. W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," Proc. of IWAENC, pp. 1–4, 2010.

[7] N. Faraji and R. C. Hendriks, "Noise Power Spectral Density estimation for public address systems in noisy reverberant environments," Proc. of IWAENC, pp. 1–4, 2012.

[8] N. Faraji, S. M. Ahadi, and H. Sheikhzadeh, "A fast semi-blind reverberation time estimation using Non-linear Least Squares Method," Lecture Notes in Computer Science: Advances in Nonlinear Speech Processing, pp. 208–215, 2013.

[9] T. de M. Prego, A. A. de Lima, S. L. Netto, B. Lee, A. Said, R. W. Schafer, and T. Kalker, "A blind algorithm for reverberation-time estimation using subband decomposition of speech signals," *Journal of the Acoustical Society of America* **131**, pp. 2811–2816, 2012.

[10] S. Broda, K. Carstensen, and M. S. Paolella, "Bias-adjusted estimation in the ARX(1) model," *Computational Statistics & Data Analysis* **51**, pp. 3355–3367, 2007.

[11] E. Eweda, "Analysis and design of a signed regressor LMS algorithm for stationary and nonstationary adaptive filtering with correlated Gaussian data," *IEEE Transactions on Circuits and Systems* **37**, pp. 1367-1374, 1990.

[12] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," 16th Int. Conf. on Digital Signal Processing, pp. 1–5, 2009.

[13] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," Proc. of ICASSP, pp. 4266–4269, 2010.