A Fast Semi-blind Reverberation Time Estimation Using Non-linear Least Squares Method

Neda Faraji, Seyed Mohammad Ahadi, and Hamid Sheikhzadeh

Speech Processing Research Lab., Amirkabir University of Technology, Tehran, Iran

Abstract. Reverberation Time (RT) estimation is of great importance in de-reverberation techniques and characterizing room acoustics. Estimating and updating the RT parameter of an enclosed environment could be carried out either continuously or discretely in the free-decaying regions of recorded reverberant signal. In this paper, we present a novel continuous sub-band-based RT estimation method which employs the general model for the Power Spectral Density (PSD) of the reverberant signal. The temporal envelope of the observed reverberant PSD in each sub-band is fitted to the temporal envelope of the proposed theoretical PSD of the reverberant signal to estimate the RT value. In Comparison to a well-known method for RT estimation, the proposed approach performs both more accurately and faster, so that it can be used in real-time applications for fast tracking of the RT value with high accuracy.

1 Introduction

Consider an enclosed environment, such as a classroom or corridor, in which a sound from a source is radiated. The receiving object, for example microphone or human being, in addition to the original radiated sound receives reverberated sounds from the surfaces in the room. In fact, the receiving microphone records the convolution of the original radiated signal with a decaying function called Room Impulse Response (RIR). Reverberation Time (RT) is an important parameter to quantify the total reverberation effect of the enclosed environment. It is defined as the time interval during which the energy of the reverberant signal decreases 60dB after playing off the radiated signal.

Identifying the RT parameter of an RIR is a challenging subject in signal processing. Some dereverberation techniques use an estimate of RT value [1][2][3]. RT estimation is also of interest for acousticians in architectural design of auditoriums and large chambers. There are some approaches for off-line measurement of the RT by radiating either a burst of noise [4] or brief pulse [5] into the test enclosure to determine the RIR. The RT can be inferred from the slope of the measured RIR. These methods require careful experiments and sufficient excitation signals. Therefore, RT estimation from a recorded reverberant signal with speech as the excitation signal is more preferable. In completely blind approaches of this category, no prior information of the room and the radiated speech is

T. Drugman and T. Dutoit (Eds.): NOLISP 2013, LNAI 7911, pp. 208–215, 2013.

available. Hence these methods can be incorporated in hearing-aids or hands-free telephony devices [6]. Generally, blind methods for RT estimation from the recorded reverberant speech can be categorized in two classes. The first class consists of locating free-decay regions in the reverberated signal, which carry more information of the RIR [1] [7] [8]. However, these methods are more vulnerable in noisy conditions as the free-decay regions have the lowest SNRs. Moreover, a large number of these free-decay parts are required for reliable estimation of RT. To overcome the need for long data recording, so that fast tracking of RT would be possible, the approach proposed in [9] utilized a sub-band decomposition to estimate RT in each sub-band of the free decay parts. In the second class, RT is estimated continuously for each arbitrary frame of the reverberant signal [6] [10]. The final RT estimate is obtained using an order-statistics filter on a number of accumulated RT estimates.

In this paper, we propose a continuous RT estimation method where the excitation signal is an available speech signal. Therefore our approach is semi-blind. We derive a theoretical model for the Power Spectral Density (PSD) of the reverberated speech [11]. The PSD of the reverberant speech depends on the RT parameter which is the desired parameter to be estimated. Through fitting the theoretical PSD to the observed reverberant PSD, RT estimation algorithm can be run segment-by-segment in each sub-band without need to seek for free-decay parts. As the theoretical PSD non-linearly depends on RT, a Non-linear Least Squares (NLS) method is employed in the estimation algorithm. Finally, statistics can be inferred from the histogram constructed based on a number of estimated RTs. Comparing different statistics, we show that the most frequently occurring estimated RT during a time interval (mode of histogram) is an accurate approximation to the real RT. We compare our continuous algorithm with a rather newly developed method [7] which estimates RT only during free-decay regions of the reverberant speech. Assuming that the offsets of speech signal occur sharply, [7] uses an approximate model which stands for free-decay parts of the reverberant speech and utilizes a Maximum Likelihood (ML) approach for RT estimation. On the other hand, our approach utilizes an exact model which stands for any arbitrary segment of the reverberant speech and therefore will be shown that outperforms [7] in accuracy. Moreover, compared to [7], our continuous approach speeds up tracking of RT.

Our proposed method can be incorporated in Public Address Systems (PAS) in which the original signal is assumed to be available. For example, [11] proposed a noise PSD estimator employed in the intelligibility improvement algorithm of a PAS, assuming reverberant enclosure. In [11], it was assumed that reverberation time of the enclosure is available. Using the RT estimation method we present in this paper, the noise PSD estimator [11] would be applicable for the environments with time-varying RT.

This paper is organized as follows. In Sec. 2, we derive a closed-form equation for the PSD of reverberated signal. Then, the proposed RT estimation algorithm and experimental results are presented in Sec. 3. Finally, we draw our conclusions in Sec. 4.

2 PSD of the Reverberated Speech

2.1 Time-Domain Model of Reverberant Speech

Assume a clean speech signal, s, is radiated through a loudspeaker in an enclosure. A microphone at a specified distance from the loudspeaker records the direct-path signal along with the reflections from the surfaces in the enclosure. We can define the observed signal at the receiver side by the following equation

$$x(l) = g(l) * s(l), \tag{1}$$

in which l is the sample index, $*$ stands for convolution operator and $g(l)$ models both the late reverberation and the direct-path as below

$$g(l) = \begin{cases} \alpha & l = 0 \\ h(l-1) & l \geq 1, \end{cases} \tag{2}$$

where h is the RIR excluding the direct path. In Polack's statistical model [12] of the RIR, a specific RIR is an ensemble of the following stochastic process

$$h(l) = b(l)e^{-\eta l} \text{ for } l \geq 0, \tag{3}$$

in which $b(l)$ is a zero-mean Normal stochastic process with variance ν^2 modulated with an exponential function with the decay rate η. The decay rate is defined as $\eta = \frac{3ln(10)}{RTf_s}$ in which RT and f_s are reverberation time and sampling frequency, respectively. Assuming that the attenuation factor α and clean speech signal s are available, we can rewrite (1) as

$$z(l) = g(l) * s(l) - \alpha s(l) = z(l) = \sum_{p=0}^{\infty} h_l(p)s(l-p-1), \tag{4}$$

in which $z(l)$ is the reverberated speech signal excluding the direct path.

2.2 Derivation of the Reverberated PSD

It has been shown [13] that

$$Z(i,k) \approx \sum_{p=0}^{\infty} h_{i+\frac{L}{2}}(p)S(i-p-1,k), \tag{5}$$

where $S(i-p-1,k)$ is the k^{th} Short Time Discrete Fourier Transform (STDFT) coefficient of a frame of clean speech starting at sample point $i-p-1$. $Z(i,k)$ is the k^{th} STDFT coefficient of the reverberated speech frame with sample index i and L shows the frame length. We derive a new equation for the PSD of reverberated speech based on (5). By definition of PSD as $\sigma_Z^2(i,k) = Var\{Z\} = E\left\{Z^2(i,k)\right\} - E^2\left\{Z(i,k)\right\}$, the first two moments of Z have to be derived.

We assume the clean signal, and therefore its spectrum, to be available. The first moment of Z is determined as follows:

$$m_Z(i,k) = E\{Z(i,k)|\mathbf{S}\} = \sum_{p=0}^{\infty} E\left\{h_{i+\frac{L}{2}}(p)S(i-p-1,k)|\mathbf{S}\right\}$$
$$= \sum_{p=0}^{\infty} E\left\{h_{i+\frac{L}{2}}(p)\right\} S(i-p-1,k) = 0, \qquad (6)$$

in which $\mathbf{S} = \{S(i-1,k), S(i-2,k), ..., S(1,k)\}$. For the second moment we have:

$$\sigma_Z^2(i,k) = E\left\{Z^2(i,k)|\mathbf{S}\right\} = E\left\{\sum_{p=0}^{\infty}\sum_{q=0}^{\infty} h_{i+\frac{L}{2}}(p)h_{i+\frac{L}{2}}(q)S(i-p-1,k)S(i-q-1,k)|\mathbf{S}\right\}$$
$$= \sum_{p=0}^{\infty}\sum_{q=0}^{\infty} E\left\{h_{i+\frac{L}{2}}(p)h_{i+\frac{L}{2}}(q)\right\} S(i-p-1,k)S(i-q-1,k). \qquad (7)$$

Refering to Polack's model (3), the expectation $E\left\{h_{i+\frac{L}{2}}(p)h_{i+\frac{L}{2}}(q)\right\} = \nu^2_{i+\frac{L}{2}} e^{-2\eta_{i+\frac{L}{2}}p}, q = p$ and $E\left\{h_{i+\frac{L}{2}}(p)h_{i+\frac{L}{2}}(q)\right\} = 0, q \neq p$. Finally, we obtain

$$\sigma_Z^2(i,k) = \sum_{p=0}^{\infty} \nu^2_{i+\frac{L}{2}} e^{-2\eta_{i+\frac{L}{2}}p} S^2(i-p-1,k). \qquad (8)$$

3 Experiments on the RT Estimation Algorithm

We tested our algorithm on six speech files of TIMIT database which were selected from 6 different speakers, 3 males and 3 females and were concatenated to construct the final test file. We generated 6 synthetic RIRs using Polack's model (3) with $RT = [0.1, 0.2, 0.4, 0.6, 0.8, 1]$ sec and $\nu^2 = 0.25$. We made six synthetically reverberated speech signals by convolving the test signal with the RIRs. In all experiments, we set the upper limit of the summation in (8) to 4000. Moreover, the clean and reverberated speech signals are segmented into 16msec frames with the frameshift of one sample at the sampling frequency of 8000 Hz. The hamming-windowed frames of both clean and reverberated speech signals are transformed into Fourier domain with 128 DFT points.

3.1 Verification of the Theoretical Model

Here, we carry out an experiment to verify the theoretically-derived reverberant PSD (8). First, for a synthetic RIR with $RT = 0.3$ sec ($\eta = 0.0028$) and $\nu^2 = 0.25$, we compute the theoretical PSD of the reverberated test speech (σ_Z^2) based

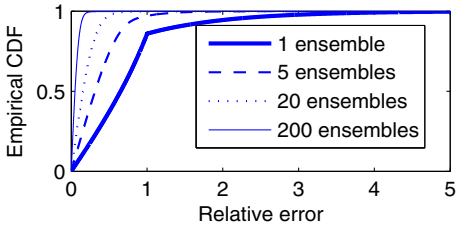

Fig. 1. Emperical CDF of the relative errors

on (8). Then, this theoretical PSD is compared with the observed PSD of the reverberant speech. We can generate an ensemble of the reverberated test speech by making a realization of the RIR with the above parameters. In application, the PSD of the observed reverberant speech is estimated by Periodogram method ($\hat{\sigma}_Z^2$), which is an approximate to the true PSD. A better approximation to the true PSD can be made by ensemble averaging of the Periodogram-based PSDs. In order to generate different ensembles of the reverberated speech, different ensembles of the RIR are realized and convolved with the test signal. Comparing the theoretical PSD with the observed PSD of the reverberated speech, the relative error is defined as follows

$$\text{Relative Error} = \left|\sigma_Z^2(i,k) - \hat{E}\left\{\hat{\sigma}_Z^2(i,k)\right\}\right| / \sigma_Z^2(i,k), \tag{9}$$

in which $\hat{E}\left\{\hat{\sigma}_Z^2(i,k)\right\}$ is the sample mean of the Periodogram-estimated PSDs of reverberant speech. The Cumulative Distribution Function (CDF) of the relative error values of all frames and frequency bins are shown in Fig. 1 with different number of ensembles to obtain the sample mean $\hat{E}\left\{\hat{\sigma}_Z^2(i,k)\right\}$. As the number of ensembles increases, better fitting between the theoretical and the ensemble-averaged PSDs is observed.

3.2 Proposed Method for RT Estimation

As shown in (8), the reverberated PSD non-linearly depends on the decay rate η. The parameter η, and therefore RT, could be simply determined through minimizing the squared error between the observed PSD and the theoretical PSD of (8). In fact, for the observed PSD of reverberant speech estimated in k^{th} frequency bin of the i^{th} frame, $\hat{\sigma}_Z^2(i,k)$, the parameter η is obtained as:

$$\hat{\eta}(i,k) = \underset{\eta}{Min}\left\{\left(\hat{\sigma}_Z^2(i,k) - \sum_{p=0}^{\infty} \nu_{i+\frac{L}{2}}^2 e^{-2\eta_{i+\frac{L}{2}}p} S^2(i-p-1,k)\right)^2\right\}. \tag{10}$$

3.3 Experimental Results of the Proposed RT Estimation Method

For each of 6 synthetically reverberated speech signals, the same experiment is carried out as follows. First, the reverberated speech is segmented into 64 ms

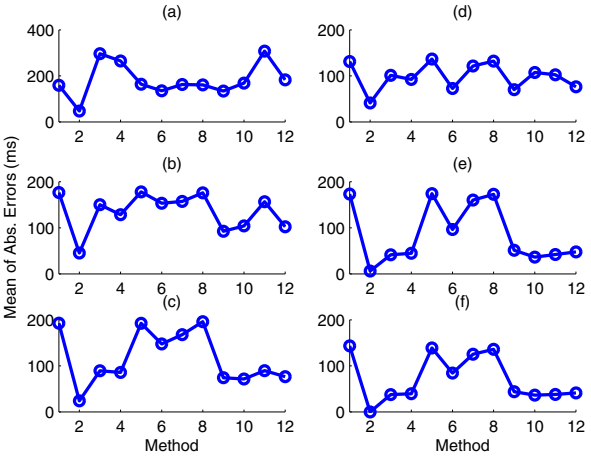

Fig. 2. Mean of Absolute Errors using different statistics (a) RT=1 s, (b) RT=800 ms, (c) RT=600 ms, (d) RT=400 ms, (e) RT=200 ms, (f) RT=100 ms

blocks with a block shift of 3ms. Then each block is divided into frames with the setup mentioned before. As mentioned, in frame scale, η can be estimated using (10). For block scale estimation, the objective function to be minimized is

$$\hat{\eta}(j,k) = \underset{\eta}{Min} \left\{ \sum_{i=1}^{M} \left(\hat{\sigma}_Z^2(i,k) - \sum_{p=0}^{4000} \nu_{i+\frac{L}{2}}^2 e^{-2\eta_{i+\frac{L}{2}} p} S^2(i-p-1,k) \right)^2 \right\}, \quad (11)$$

where j is the block index and i represents the frame number in the block. For our setup, M is set to 384. To infer the fullband η from the subband block-based estimated η parameters, different statistics could be employed. For example, similar to the method proposed in [9], we employ the Median of Medians method. First, the fullband RT of the j^{th}block is set to the median of the RTs derived in all subbands of the block. Then, the fullband RT of the total reverberant speech is estimated through computing median of all J fullband block-derived RTs

$$\hat{RT} = \underset{j=1,2,\ldots,J}{Median}\, \hat{RT}(j) = \underset{j=1,2,\ldots,J}{Median} \left\{ \underset{k=1,2,\ldots,NFFT/2+1}{Median}\, \hat{RT}(j,k) \right\}. \quad (12)$$

In another approach [7][1] the mode of the histogram of the block-estimated RTs was considered as the final RT estimate. Hence, in our proposed approach we compare different statistics together. First, we create the histogram of all ηs derived from all frequency bands of the N consecutive blocks. We measure four statistics from this histogram that are Median, Mode, Mean and Mean after removing 30 percent of outliers (Trimmed Mean). The inferred statistics constitute

[1] Implemented code is available at `http://www.mathworks.com/matlabcentral/fileexchange/35740-blind-reverberation-time-estimation`

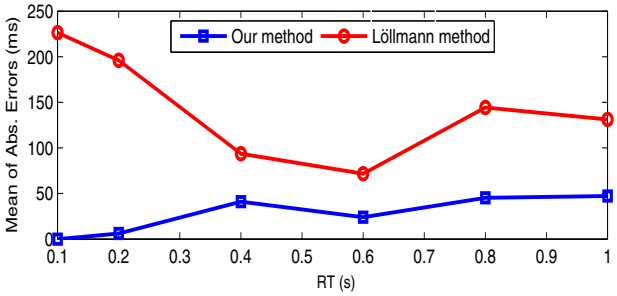

Fig. 3. Average of Means of Abs. Errors using method 2 and Löllmann method [7].

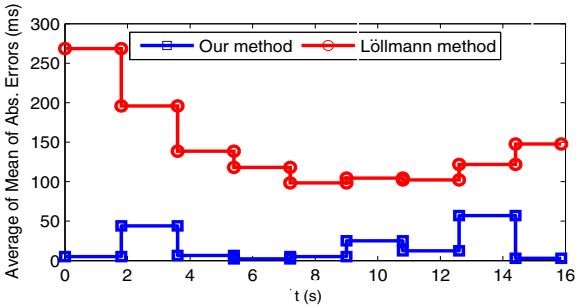

Fig. 4. Trend of the Average of Means of Absolute Errors in the intervals of 1.8 sec

Methods 1 to 4 in the depicted results in Fig. 2. Also, we can extract the Median value of $NFFT/2+1$ estimated ηs in each block and then make the histogram for the N Median values. Hence, the other four statistics 5 to 8 are Median of Medians, Mode of Medians, Mean of Medians and Trimmed Mean of Medians. The same idea can be used for the Mean value of the ηs extracted in each block. Therefore, methods 9 to 12 involve Median of Means, Mode of Means, Mean of Means and Trimmed Mean of Means. In our setup, we used 300 blocks to make the histogram, which for the block shift of 3msec corresponds to the time interval of 900 msec. Then, similar to [7], we smooth η using a recursive averaging filter with the time-constant of 0.996. Finally, we obtain the estimated RT as $\hat{RT} = \frac{3\times ln(10)}{\hat{\eta} f_s}$. The performance can be quantified by averaging the Absolute Errors between the target RT and the estimated RTs as below

$$\text{Mean of Absolute Errors} = \frac{1}{J}\sum_{j=1}^{J}\left|\hat{RT}(j) - RT\right|. \tag{13}$$

In Fig. 2, Means of Absolute Errors for six reverberated speech signals are demonstrated. The horizontal axis represents different statistics. As shown, method 2 has the minimum Mean of Absolute Errors compared to the other methods. Using method 2, Means of Absolute Errors for 6 reverberated signals are plotted in Fig. 3, in which the performance of the method of Löllmann et al. [7] is depicted

too. Besides, in real situations, the reverberation time varies with time, so that it is of advantage to have a fast RT estimation algorithm. As our procedure is able to extract the RT for any arbitrary segment of the reverberated speech, it is considerably faster compared to [7] that estimates the RT in the free decay parts of the reverberated speech. In order to illustrate the high speed of our algorithm, for each of 6 reverberant speech signals, we compute the Mean of Absolute Errors in the interval of 1.8 sec rather than the total reverberated speech. The average of the short-time Means of Absolute Errors over the 6 reverberant signals is computed. Fig. 4 demonstrates the trend of this average over time. Compared to [7], our algorithm is able to detect the RT even in short segments of the reverberated speech.

4 Conclusion

In this paper we have proposed a continuous RT estimation algorithm based on a general model we derived for PSD of reverberant speech. The presented algorithm works on subband domain to extract the RT of any arbitrary segment of the reverberant speech. Compared with a new method of Löllmann et al., our approach achieves superior performance with fast adaptation speed.

References

1. Lebart, K., Boucher, J.M.: A new method based on spectral subtraction for speech dereverberation. Acta Acoustica-ACOUSTICA 87, 359–366 (2001)
2. Habets, E.A.P.: Single-channel speech dereverberation based on spectral subtraction. In: Proc. of Ann. Workshop on Circuits, Systems and Signal Processing (2004)
3. Jan, T., Wang, W.: Joint blind dereverberation and separation of speech mixtures. In: Proc. of EUSIPCO, pp. 2343–2347 (2012)
4. ISO-3382, Acoustics measurement of the reverberation time of rooms with reference to other acoustical parameters, Int. Org. for Standardization, Geneva (1997)
5. Schroeder, M.R.: New method for measuring reverberation time. Journal of the Acoustical Society of America 37, 409–412 (1965)
6. Ratnam, R., Jones, D.L., Wheeler, B.C., O'Brien, W.D., Lansing, C.R., Feng, A.S.: Blind estimation of reverberation time. Journal of the Acoustical Society of America 114, 2877–2892 (2003)
7. Löllmann, H.W., Yilmaz, E., Jeub, M., Vary, P.: An improved algorithm for blind reverberation time estimation. In: Proc. of IWAENC (2010)
8. Vesa, S., Harma, A.: Automatic estimation of reverberation time from binaural signals. In: Proc. of ICASSP, pp. 281–284 (2005)
9. de, T., Prego, M., de Lima, A.A., Netto, S.L., Lee, B., Said, A., Schafer, R.W., Kalker, T.: A blind algorithm for reverberation-time estimation using subband decomposition of speech signals. Journal of the Acoustical Society of America (2012)
10. Wen, J.Y.C., Habets, E.A.P., Naylor, P.A.: Blind estimation of reverberation time based on the distribution of signal decay rates. In: Proc. of ICASSP (2008)
11. Faraji, N., Hendriks, R.C.: Noise Power Spectral Density estimation for public address systems in noisy reverberant environments. In: Proc. of IWAENC (2012)
12. Polack, J.D.: La transmission de l' energie sonore dans les salles (1988)
13. Erkelens, J.S., Heusdens, R.: Correlation-based and model-based blind single channel late-reverberation suppression in noisy time-varying acoustical environments. IEEE Trans. Audio, Speech and Language Processing 18, 1746–1765 (2010)