PERFORMANCE COMPARISON OF ALGORITHMS FOR BLIND REVERBERATION TIME ESTIMATION FROM SPEECH

N. D. Gaubitch[1]*, H. W. Löllmann*[2]*, M. Jeub*[2]*, T. H. Falk*[3]*, P. A. Naylor*[1]*, P. Vary*[2]*, M. Brookes*[1]

[1]Centre for Law Enforcement Audio Research (CLEAR), Imperial College London, UK
[2]Inst. of Comm. Systems and Data Proc., RWTH Aachen University, Germany
[3]Institute National de la Recherche Scientifique, Montreal, Canada

ABSTRACT

The reverberation time, T_{60}, is one of the key parameters used to quantify room acoustics. It can provide information about the quality and intelligibility of speech recorded in a reverberant environment, and it can be used to increase robustness to reverberation of speech processing algorithms. T_{60} can be determined directly from a measurement of the acoustic impulse response, but in situations where this is unavailable it must be estimated blindly from reverberant speech. In this contribution, we provide a study of three state-of-the-art methods for blind T_{60} estimation. Experimental results with a large number of talkers, simulated and measured acoustic impulse responses, and various levels of additive white Gaussian noise are presented. The relative merits of the three methods in terms of computational time, estimation accuracy, noise sensitivity and inter-talker variance are discussed. In general, all three methods are able to estimate the reverberation time to within 0.2 s for $T_{60} \leq 0.8$ s and SNR ≥ 30 dB, while increasing the noise level causes overestimation. The relative computational speed of the three methods is also assessed.

Index Terms— Reverberation time, blind estimation

1. INTRODUCTION

When a speech signal, $s(n)$, is produced at a point in a room it follows multiple paths to any observation point – the direct path and multiple reflections off the walls and objects in the room. This results in a reverberant observation, $x(n)$, characterized by the Acoustic Impulse Response (AIR), $h(n)$, which is a function of the room geometry, the reflectivity of the walls and other objects, and the source-microphone distance and location. The observed signal is the convolution between the AIR and the original speech signal and, inevitably, with additive measurement noise, $\nu(n)$

$$x(n) = s(n) * h(n) + \nu(n). \quad (1)$$

Reverberation time is one of the key parameters used to quantify room acoustics. It follows the pioneering work of W. C. Sabine in the 19th century where he found that a sound source becomes inaudible when it has decayed by 60 dB after becoming inactive. Consequently, reverberation time is defined as the time it takes for the sound to decay by 60 dB once the source has been switched off [1] and is denoted here by T_{60}. The parameter T_{60} is a function of the room geometry and the reflectivity of the surfaces in the room; this relationship is commonly given by the Sabine or the Eyring equations [2, 3]. As opposed to the AIR, T_{60} is independent of the source-microphone configuration. Most often, T_{60} is given as a single value but could also be given for different frequency bands, for example, octave or 1/3-octave bands. An estimate of the reverberation time of a particular room can serve as an indicator of the quality and the intelligibility of speech observed in that room. It can also be used to improve the performance of speech processing applications such as speech recognition [4] and dereverberation [5, 6, 7, 3].

There are several standardized methods for estimating T_{60} from a measured AIR [8]. The most commonly used method calculates the Energy Decay Curve (EDC) using the Schroeder backward integration method [9] and fits a line to its slope in some range depending on the estimated noise floor, typically between -5 and -35 dB. The T_{60} is then estimated from the slope of the line. Although, this provides accurate estimates of T_{60}, it may not always be practical or even possible to measure the AIR in a room. Therefore, it is desirable to be able to estimate the T_{60} from an observed reverberant speech signal, $x(n)$, directly. Several algorithms have been proposed for such blind T_{60} estimation [10, 11, 12, 13, 14, 15].

In this paper, we investigate three state-of-the art methods for blind estimation of T_{60}, which are summarized in Section 2. All three methods have been shown to provide accurate estimates of T_{60} on different data sets. The objective here is to perform a comparative set of experiments using the same evaluation methodology and data, as described in Section 3, in order to investigate these methods' performance in both noisy and noise-free conditions. In particular, we look at estimation accuracy and variance with different talkers and different amounts of reverberation and the computational efficiency of the methods. The performance results and the relative merits of the different methods are discussed in Section 4 and conclusions are presented in Section 5.

2. METHODS FOR BLIND T_{60} ESTIMATION

In this section, we provide a description of the three algorithms that were investigated.

2.1. Method 1: Spectral Decay Distributions (SDD)

The method proposed by Wen *et al.* [12] is based on the spectral decay distributions of the observed speech and assumes a statistical model for the AIR. Frequency dependent decay rates are estimated by applying a least squares linear fit to the log-energy envelope in each frequency band in the Discrete Fourier Transform (DFT) domain. The negative-side variance of the distribution of the decay rates is demonstrated to correlate with the room decay rate and is, thus, used to predict reverberation time. This approach requires training in order to map the values from the negative-side variance to T_{60}. In the training phase, a 2^{nd}-order polynomial mapping function is calculated using reverberated speech with known T_{60}.

2.2. Method 2: Modulation Energy Ratios (MER)

Falk and Chan [13] proposed a non-intrusive quality measure for dereverberated speech based on the Speech-to-Reverberation Modulation energy Ratio (SRMR). The method considers the energy in eight modulation frequency bands, varying logarithmically between 4 and 128 Hz and calculated from 23 acoustic frequency bands obtained from a gammatone filterbank. It is observed that the low modulation frequency energy ($4-18$ Hz) is relatively insensitive to reverberation while the energy at high modulation frequencies ($29-128$ Hz) increases almost linearly with T_{60}. This leads to the SRMR measure which is the ratio of the average energy in the low modulation frequencies to the high modulation frequencies. It is also shown that the inverse of the SRMR is highly correlated with T_{60}. Obtaining the values for T_{60}, requires some form of training and, similarly to Method 1, a 2^{nd}-order polynomial mapping function is calculated from reverberant speech with known T_{60}.

2.3. Method 3: Maximum Likelihood (ML)

The method proposed by Löllmann *et al.* [11] is inspired by the method from Ratnam *et al.* [10]. It uses a statistical model of the sound decay of reverberant speech, following a reverberation model similar to that of Method 1. This is then used to develop a Maximum Likelihood (ML) approach for the T_{60} estimation. In order to improve the computational efficiency, the speech signals are downsampled before the estimation and there is a pre-selection approach to detect plausible decays before these are used in the ML estimation procedure. Furthermore, the estimated T_{60} for each frame is used in a histogram and smoothing procedure in order to increase the robustness of the estimates. This algorithm has also an option for a fast tracking of the T_{60}. However, tracking is not considered in this evaluation. Unlike the previous two methods, this method does not require training.

3. EXPERIMENTAL SETUP

The objective with the experiments was to investigate the estimation accuracy and the computational efficiency of the three methods described in Section 2. The following two metrics were used for evaluation:

1. *Estimation error*, defined as the difference between the estimated reverberation time, $\hat{T}_{60}$, and the true reverberation time, T_{60}
$$E = \hat{T}_{60} - T_{60} \text{ s}, \tag{2}$$
and in particular the distribution of these errors with different talkers and different levels of noise and reverberation. Positive and negative estimation errors indicate over- and underestimation, respectively.
2. *Real-time factor*, defined as the ratio between the time taken to process a sentence, T_p, and the length of the sentence, L_s
$$R = T_p/L_s. \tag{3}$$

For our experiments, the processing time was measured as the execution time in Matlab for each of the methods using the `tic` and `toc` operations. Matlab implementations of the methods were provided by their respective authors.

Anechoic speech from the TIMIT corpus [16] was used for all experiments. TIMIT contains ten sentences spoken by each of the 438 male and 192 female talkers, giving a total of 6300 sentences. The data is divided into two mutually exclusive sets: a training set and a test set. Speech from the training set was used to calculate mapping functions for Method 1 and Method 2 while the test set was used for evaluation. The sentences for each talker were concatenated to form utterances of approximately 30 s. For the test set, this resulted in 168 utterances. The sampling frequency was set to $f_s = 8$ kHz.

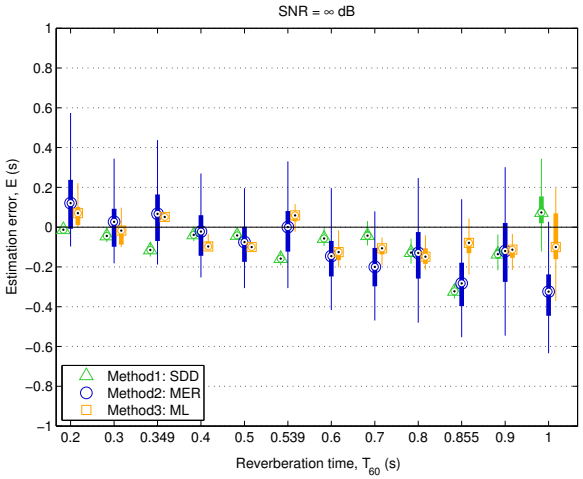

Fig. 1: Reverberation time estimation in noise-free reverberant speech. The groups of box plots above each T_{60} show the distribution of errors for the three methods. The markers indicate the method and the dots inside the markers denote the median of the estimation errors; thick vertical lines show the upper and lower quartiles and thin vertical lines indicate the estimation errors up to 1.5 times the interquartile range, covering approximately 99.3% of the data. $T_{60} = \{0.349, 0.539, 0.855\}$ represent measured AIRs.

Three sets of AIRs were considered. The first set consists of simulated impulse responses using the source-image method [17] for a room with dimensions $(6 \times 5 \times 4)$ m and a source-microphone distance of 1.5 m. The reverberation time varies between 0.2 and 1 s in steps of 0.1 s. The second set uses three AIRs from the Aachen Impulse Response database [18] measured in an office, a meeting room, and a lecture room with microphone-source distance of 2 m (measurements excluding dummy head). The reverberation times were calculated using the $1/3$-octave band procedure in the ISO3382 standard [8] resulting in 0.349 s, 0.539 s and 0.855 s for the office, the meeting room, and the lecture room, respectively; these were used as ground truth T_{60} values in the evaluation. Finally, the third set comprises simulated AIRs for T_{60} between 0.1 and 1.1 s in steps of 0.05 s using Polack's statistical model [5, 6].

Following the signal definition in (1), reverberant and noisy speech was generated by first convolving the anechoic speech with the AIRs and then adding White Gaussian Noise (WGN) to the reverberated speech at Signal-to-Noise Ratios (SNRs) of $0, 20, 40, \infty$ dB, where SNR $= \infty$ represents the noise-free case. The noisy sample was generated by calculating the active speech level in the reverberant speech using the method in ITU-T P.56 [19] and adjusting the noise level to the desired SNR.

The simulated impulse responses from the Polack's method were used with sentences from the training set of TIMIT for the calculation of the mapping functions for Methods 1 and 2. In this way, different speech and impulse responses were used for training and testing. Furthermore, none of the methods had noise-robustness included in the training – all training samples were at SNR $= \infty$ dB.

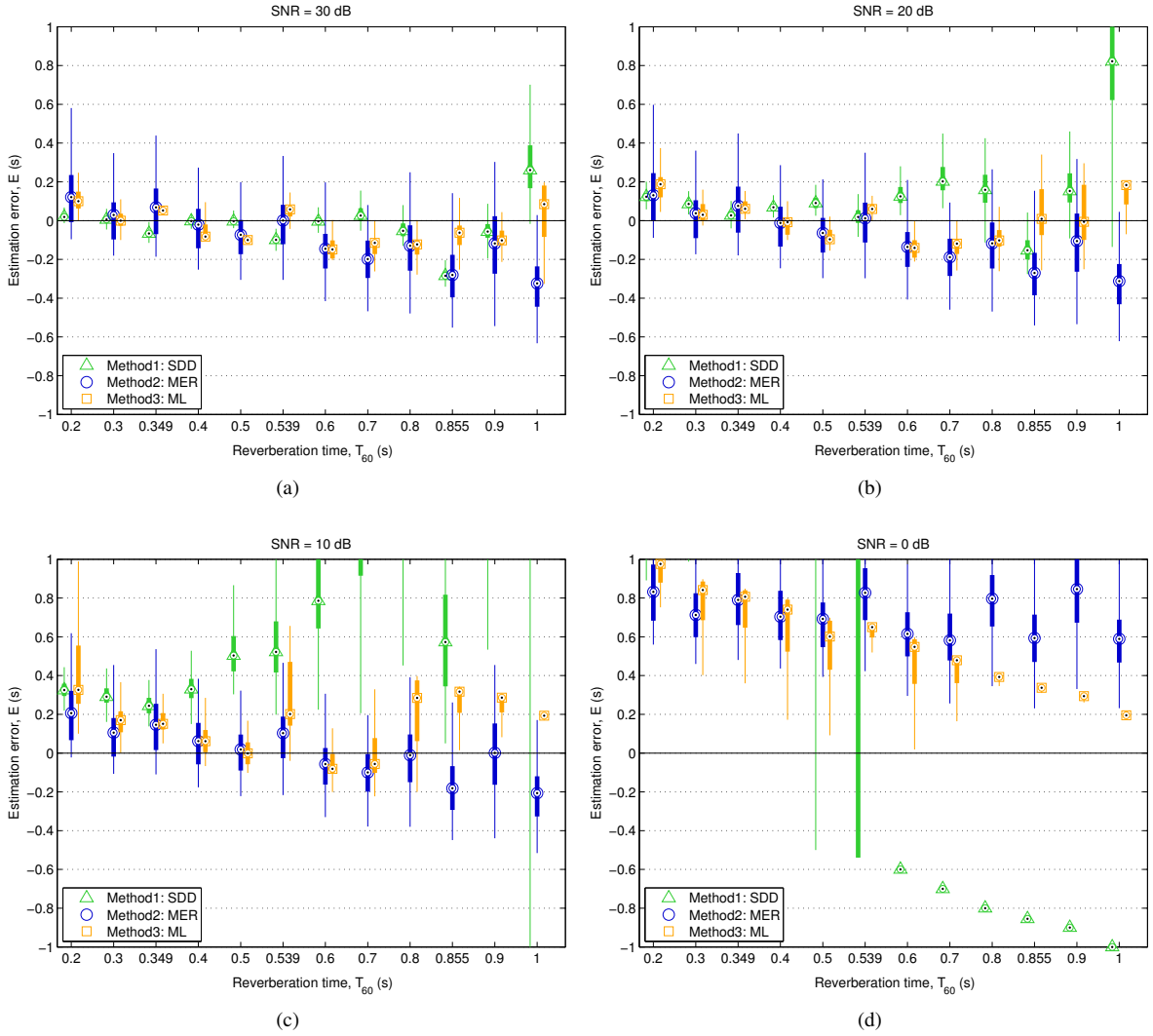

Fig. 2: Reverberation time estimation in additive white Gaussian noise at (a) SNR = 30 dB, (b) SNR = 20 dB, (c) SNR = 10 dB and (d) SNR = 0 dB. See Fig. 1 for explanation of the plots.

4. RESULTS AND DISCUSSION

We now present the results for the three methods evaluated according to the experimental setup in Section 3. Figure 1 shows the results for the noise-free case, SNR $= \infty$ dB, and Figs. 2a-2d show the results for SNR ranging from 30 dB to 0 dB, respectively. The results for both simulated and measured AIRs are shown on the same box plots of the estimation errors, E versus increasing reverberation time; $T_{60} = \{0.349, 0.539, 0.855\}$ represent measured AIRs from [18]. Markers indicate the different algorithms: Method 1 (triangles), Method 2 (circles) and Method 3 (squares). The dots inside the markers are positioned at the median of the estimation errors, the thick vertical lines represent the lower and the upper quartiles and the thin vertical lines represent the remaining data points up to ± 1.5 times the inter-quartile range giving approximately 99.3% data coverage. We show only up to ± 1 s of estimation error for size reasons.

We see from Fig. 1 that Methods 1 and 3 provide accurate estimates to within ± 0.2 s for all cases of $T_{60} \leq 0.8$ with little inter-talker variance, after which the inter-talker variance increases. Although, on average, Method 2 results in similar estimation accuracy within that range, it exhibits a larger iner-talker variation.

The two main effects of additive noise on all three methods is a positive bias, i.e. overestimation, and an increased inter-talker variance of the estimation error. Despite the larger inter-talker variance, Method 2 exhibits the lowest sensitivity to noise and results in similar performance for all SNR ≥ 10 dB. At SNR $= 0$ dB Methods 1 and 2 fail to provide estimates within the ± 0.5 s range, while Method 3 is able to do so for $T_{60} \geq 0.7$.

Table 1 lists the mean, μ_R, and standard deviation, σ_R, of the real time factors averaged over the estimations for all talkers, reverberation times and SNRs. It can be seen that Methods 2 and 3 are both able to operate in real time in our Matlab implementation ($R < 1.0$), while Method 1 is not. The very low real-time factor for Method 3 is evidence that this algorithm has been successfully designed for real-time speech dereverberation [7].

The seemingly similar effect of noise on all three methods comes

	Real-time factor $\mu_R \pm \sigma_R$
Method 1: SDD	2.066 ± 0.153
Method 2: MER	0.284 ± 0.069
Method 3: ML	0.062 ± 0.017

Table 1: Real-time factor for each of the three methods as an average over all 168 utterances.

from the fact that they are all related in the sense that the estimates of T_{60} depend, to a greater or lesser extent, on the modulations of the reverberant speech. As the level of additive noise increases, the modulations are buried in it and will tend to the modulation spectrum of the noise – a fact that has previously been exploited to study the effects of noise and reverberation on speech intelligibility [20]. This will have the greatest effect on Methods 1 and 3 where the estimated decays will be biased positively by additive noise. On the other hand, Method 2 is more resilient to noise since it uses the ratio of the modulation frequencies, which would be less affected by white noise, in particular. Since Methods 1 and 2 use mapping functions, these results suggest that noise robustness could be added by training at different SNRs and using an estimate of the SNR as a parameter, as shown in, for example, [14].

5. CONCLUSIONS

We presented a comparative quantitative study of three methods for blind estimation of reverberation time: a method based on speech slope distributions, a method based on modulation energy ratios and a method based on a maximum likelihood estimation of the reverberation tail slope. The investigation considered estimation accuracy, talker dependency and relative computation time. Experiments were performed on a variety of measured and simulated AIRs and different levels of additive white Gaussian noise. The results showed that the methods based on speech slope distributions and maximum likelihood provide accurate estimates to within ± 0.2 s for $T_{60} \leq 0.8$ s for SNRs greater than 30 dB. The modulation energy ratio based method exhibits a larger inter-talker variance but is less sensitive to noise, providing similar performance down to $SNR = 10$ dB.The key effect of additive noise on all methods is a positive bias of the estimation error which results in estimation errors of $E \geq 0.2$ s for all methods. The speech slope distribution based method was the most demanding computationally, while the remaining two are must faster with the maximum likelihood algorithm having a real time factor of 0.062 in terms of Matlab execution time in the implementation used in our tests.

6. REFERENCES

[1] W. C. Sabine, *Collected Papers on acoustics (Originally 1921)*, Peninsula Publishing, 1993.

[2] Carl F. Eyring, "Reverberation time in "dead" rooms," *J. Acoust. Soc. Am.*, vol. 1, no. 2, pp. 217–241, Jan. 1930.

[3] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer, 2010.

[4] L. Couvreur and C. Couvreur, "On the use of artificial reverberation for ASR in highly reverberant environments," in *Proc. 2nd IEEE Benelux Signal Processing Symposium (SPS-2000)*, Hilvarenbeek, The Netherlands, Mar. 2000, pp. S001 – S004.

[5] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

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

[7] M. Jeub, H. W. Löllmann, , and P. Vary, "Blind dereverberation for hearing aids with binaural link," in *Proc. ITG Conf. Speech Comm.*, Bochum, Germany, Oct. 2010.

[8] ISO3382, "Acoustics – measurement of room acoustic parameters," 2009.

[9] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Am.*, vol. 37, pp. 409–412, 1965.

[10] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien, Jr., C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *J. Acoust. Soc. Am.*, vol. 114, no. 5, pp. 2877–2892, Nov. 2003.

[11] H. W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Tel-Aviv, Israel, Aug. 2010.

[12] J. Y. C. Wen, E. A. P. Habets, and P. A. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, USA, Apr. 2008.

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

[14] T. H. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 978–989, Apr. 2010.

[15] T. de M. Prego, A. A. de Lima, S. L. Netto, B. Lee, R. W. Schafer, and T. Kalker, "A blind algorithm for reverberation-time estimation using subband decomposition of speech signals," *J. Acoust. Soc. Am.*, vol. 131, no. 4, pp. 2811–2816, Apr. 2012.

[16] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," Technical report, National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Dec. 1988.

[17] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[18] M. Jeub, M. Schäfer, and P. Vary, "A binaural impulse response database for evaluation of derverberation algorithms," in *Proc. Intl. Conf. Digital Signal Process.*, Santorini, Greece, July 2009.

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

[20] F. Dubbelboer and T. Houtgast, "The concept of signal-to-noise ratio in the modulation domain and speech intelligibility," *J. Acoust. Soc. Am.*, vol. 124, no. 6, pp. 3937–3946, Dec. 2008.