

BLIND ESTIMATION OF REVERBERATION TIME BASED ON THE DISTRIBUTION OF SIGNAL DECAY RATES

Jimi Y.C. Wen¹, Emanuël A.P. Habets² and Patrick A. Naylor¹

¹ Department of EEE
Imperial College, London, UK
{jung.wen,p.naylor}@imperial.ac.uk

² School of Engineering
Bar-Ilan University, Ramat-Gan, Israel
habetse@eng.biu.ac.il

ABSTRACT

The reverberation time is one of the most prominent acoustic characteristics of an enclosure. Its value can be used to predict speech intelligibility, and is used by speech enhancement techniques to suppress reverberation. The reverberation time is usually obtained by analysing the decay rate of i) the energy decay curve that is observed when a noise source is switched off, and ii) the energy decay curve of the room impulse response. Estimating the reverberation time using only the observed reverberant speech signal, i.e., blind estimation, is required for speech evaluation and enhancement techniques. Recently, (semi) blind methods have been developed. Unfortunately, these methods are not very accurate when the source consists of a human speaker, and unnatural speech pauses are required to detect and/or track the decay. In this paper we extract and analyse the decay rate of the energy envelope blindly from the observed reverberation speech signal in the short-time Fourier transform domain. We develop a method to estimate the reverberation time using a property of the distribution of the decay rates. Experimental results using simulated and real reverberant speech signals demonstrate the performance of the new method.

Index Terms— reverberation time, blind estimation, acoustic signal analysis.

1. INTRODUCTION

The problem of reverberation is important for both the audio signal processing and room acoustics community. Reverberation is caused by the multi-path propagation of acoustic signals from a source to a microphone. Reverberant speech can be described as sounding distant with noticeable echo and colouration. The human auditory system is believed to have echo suppression and dereverberation capabilities, which are not present when sound is captured by microphones, such as in hands-free telecommunication devices. The characteristics of reverberation [1] can be derived from the room impulse response (RIR), such as *reverberation time* (RT), *definition* (Deutlichkeit), *clarity index*, and the *centre time*. There are also signal dependent approaches, e.g., the *modulation transfer function* (MTF) and the *speech transmission index* (STI). In particular, the reverberation time is still considered as the objective quantity in room acoustics.

In the early 20th century, Sabine [1] provided an empirical formula to predict the RT in an enclosure. The formula is based solely on the geometry and the surface material of the environment. Other methods measure the RT by analysing the decay rate of the sound decay curve. The decay curve can be observed when an excitation signal is switched off after reaching a steady-state sound level in

the enclosure. This method is also known as the Interrupted Noise Method (ISO 3382) [2]. Schroeder developed a method [3] to calculate the ensemble average of the decay curves directly using backwards integration of the related RIR.

Semi-blind methods have been developed, where the characteristics of the enclosures are learned using neural network approaches [4]. Another method is segments speech to detect gaps in sound so as to allow the sound decay curve to be tracked [5, 6]. An essential tool for the study of reverberation is a method to estimate reverberation characteristics from the microphone signal alone, such as that proposed in [7]. In [7], Ratnam develops a truly blind method for estimating the RT using a maximum-likelihood procedure. The estimates are obtained continuously and an ordered statistics filters is used to extract the most likely RT from the accumulated estimates [7]. To reliably extract the RT, this method requires long pauses in the speech utterance.

In this paper we develop a novel blind RT estimation method that takes into account the interaction between the decay rates of the room and speech. The estimator is based on a time-frequency room decay model which is related to Polack's statistical reverberation model [8]. A least squares method is used to continuously estimate the decay rate of the received signal in the short-time Fourier transform (STFT) domain. The time-frequency analysis is advantageous in two ways. Firstly, the requirement for long speech pauses is removed since it is sufficient to have any endpoints of the signal only over the bandwidth of the frequency bin in question. Secondly, since reverberation is frequency dependent, obtaining an estimate of the decay rate for each frequency bin can be advantageous for frequency domain enhancements [5, 9] and evaluation [10] methods. The RT is then extracted from a property of the distribution of the reverberant speech decay rates.

2. ROOM DECAY MODEL

Reverberation, described by the RIR, consists of a direct sound and early reverberation followed by late reverberation. While the fine structure of late reverberation can be modeled statistically, the decaying envelope of the RIR can be modeled as a deterministic signal parameterized by some damping constant, δ [1, 8]. Polack developed a time-domain model that describes a RIR as one realization of a non-stationary stochastic process [8]:

$$h(t) = b(t)e^{-\delta t} \quad \text{for } t \geq 0, \quad (1)$$

where $b(t)$ is a centered stationary Gaussian noise, and damping constant, δ is related to the reverberation time, RT by:

$$RT = 3 \ln 10 / \delta. \quad (2)$$

It should be noted that the relation between the damping constant δ and the RT is only valid when the sound field in the enclosure is diffuse and the source-microphone distance is greater than the critical distance [1]. The room decay model can be defined using (1) as:

$$\mathcal{E}\{h^2(t)\} = \sigma_b^2 e^{-2\delta t} = \sigma_b^2 e^{\lambda_h t}, \quad (3)$$

where σ_b^2 denotes the variance of $b(t)$, and the decay rate, $\lambda_h = -2\delta$. The room decay can be extended for frequency dependent decay rates by rewriting (1) as:

$$\tilde{H}(t, f) = P(f)e^{\lambda_h(f)t} \quad \text{for } t \geq 0, \quad (4)$$

where $\tilde{H}(t, f)$ is the energy envelope of RIR at time t and frequency f , $\lambda_h(f)$ is the decay rate at frequency f , and $P(f)$ is the initial power spectral density. The frequency dependent room decay model (4) can be linearized by taking the natural logarithm:

$$\ln \tilde{H}(t, f) = \ln P(f) + \lambda_h(f)t \quad \text{for } t \geq 0. \quad (5)$$

The decay rate $\lambda_h(f)$ can therefore be estimated by applying a linear fit to the natural logarithm of the time-frequency energy envelope.

3. ANECHOIC AND REVERBERANT DECAY RATES

In this section we analyse the decay rate of the room, as well as that obtained from anechoic and reverberant speech signal.

A linear least squares fit is applied to the natural logarithm of the time-frequency envelopes to estimate the frequency dependent decay rate, $\lambda(f)$. We note that fitting in the STFT domain has an effect of smoothing of the fine structure of the stochastic decay. Since our discussion applies to both fullband and subband signals, the frequency index f has been omitted for simplicity of notation. Furthermore, in the sequel we assume that the frequency bins are mutually independent.

Reverberant speech can be modeled as the convolution of the anechoic speech signal and the RIR. Let us assume that the anechoic signal and RIR consists of mutually uncorrelated white noise sequences, with energy envelopes $d_s(t)$ and $d_h(t)$, respectively. The energy envelope of the reverberant speech signal, denoted by $d_x(t)$, can be written as [11]:

$$d_x(t) = d_h(t) * d_s(t). \quad (6)$$

We now make two important observations: i) The speech pauses do not have instantaneous onsets because the source speech signal often decays smoothly to zero depending on the phonetic context. We denote this the speech endpoint decay. ii) The room response can be sensed when the source speech signal is zero. Combining these two observations, we note that the signal measured by a microphone during pauses in the source speech signal contains the result of convolution of the speech endpoint decay with the room decay. If the source speech signal contains any instantaneous endpoints, the speech pauses would contain the room decay. After a speech endpoint, the energy envelope of the reverberant signal can be expressed as

$$d_x(t) = e^{\lambda_h t} * e^{\lambda_s t} = \begin{cases} (e^{\lambda_h t} - e^{\lambda_s t})/(\lambda_h - \lambda_s) & \text{for } \lambda_h \neq \lambda_s \\ t e^{\lambda_h t} & \text{if } \lambda_h = \lambda_s. \end{cases} \quad (7)$$

where λ_h and λ_s denote the decay rates of the room and anechoic speech, respectively. The sum of two exponential terms will be dominated by the exponential term with the largest value. Note that (7)

can also be used to describe the energy envelope at time instances other than the speech endpoint, such as a speech onset. Therefore, the decay rate λ_x can be approximated as:

$$\lambda_x \approx \max[\lambda_h, \lambda_s]. \quad (8)$$

This approximation becomes more accurate when $|\lambda_h - \lambda_s|$ is large. In Section 4, this relation is used to explain the distribution of the reverberant speech decay rate λ_x .

4. DECAY RATE DISTRIBUTION

In this section we study the distribution of the estimated decay rates obtained from the energy envelope of the room, the anechoic and the reverberant speech signals. In the sequel the distributions are plotted by superimposing all frequency dependent decay rates.

The estimated decay rates of the RIR are dominated by the decaying energy envelope of the RIR, e.g. the true decay rate λ_h . Therefore, the mean of the estimates should be that of the decay rate of the envelope. However, there are errors in the estimated decay rates due to the of random nature of the fine structure of the RIR, which is not completely smoothed by the fitting process in the STFT domain. Fig. 1 (a-d) shows the distribution of the estimated decay rates of four RIRs with different reverberation times. It can be seen that the mean of the estimated decay rates corresponds very well to the true decay rate λ_h . The variance depends on decay estimation process, e.g., the number of frames that is used in the Least Squares (LS) fitting process and the parameters of the STFT and seem to be equivalent in each of the four cases.

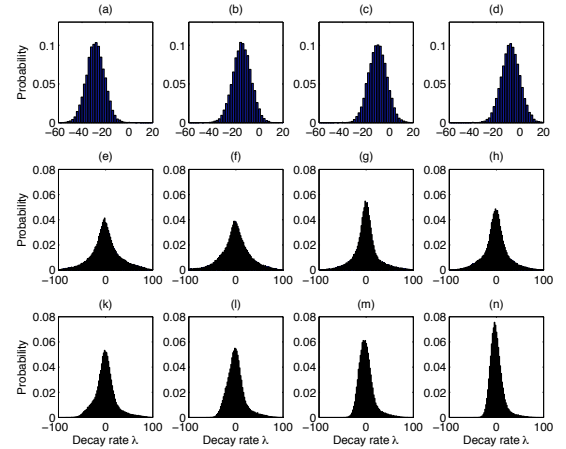


Fig. 1. Distribution of the room decay rates for all frequencies for a given reverberation time RT of (a) [$\lambda_h=-27$, $RT=250$ ms] (b) [$\lambda_h=-14$, $RT=500$ ms] (c) [$\lambda_h=-9$, $RT=750$ ms] (d) [$\lambda_h=-7$, $RT=1000$ ms]. Distribution of the speech decay rates for (e) male speaker: utterance 1, (f) utterance 2; (g) female speaker: utterance 1, (h) utterance 2. Distribution of the reverberant speech decay rates obtained using the RIRs (k)(l)(m)(n).

The envelope of the speech signal, unlike the RIR, has a non-constant decaying envelope due to the nature of speech. Many studies have been performed to find a reasonable model for the probability density function of the spectral coefficients of the speech signal. However, it is out of the scope of this paper to study the distribution of the speech decay rate for a given spectral distribution of speech, i.e., the distribution of the speech decay rate is simply observed. Fig. 1 shows the distribution of four speech fragments, one

male speaker of two utterances, Fig. 1(e) and (f); one female speaker of two utterance, Fig. 1(g) and (h). In Figure 1(k), (l), (m) and (n), the distributions of the decay rate of the reverberant speech signals are shown. It can be seen that the distribution is ‘skewed’ more as the decay rate tend to zero (or infinite RT).

4.1. Reverberant Decay Rate Distribution

In this part we analyse the distribution of reverberant speech decay rate. The distribution of the reverberant speech decay rate, $f_x(\lambda)$, can be written as a function of the speech decay rate distribution, $f_s(\lambda)$, and the room decay rate distribution, $f_h(\lambda)$:

$$f_x(\lambda) = \int_{-\infty}^{\infty} g(f_s(\tau), f_h(\lambda)) d\tau. \quad (9)$$

Using the approximation in (8), and the fact that λ_s and λ_h are independent, the function $g(f_s(\tau), f_h(\lambda))$ can be written as:

$$g(f_s(\tau), f_h(\lambda)) = \begin{cases} f_s(\tau) f_h(\lambda) & \text{if } \lambda > \tau \\ f_s(\tau) F_h(\tau) & \text{if } \lambda = \tau \\ 0 & \text{otherwise,} \end{cases} \quad (10)$$

where $F_h(\lambda)$ is the cumulative distribution function of $f_h(\lambda)$. The distribution of the reverberant speech decay rates $f_x(\lambda)$ can be thought of as the sum of infinitely many processes of g between the corresponding *partial* speech decay rate distributions $f_s(\tau)$ with infinitesimal width, and the *whole* room decay rate distribution, $f_h(\lambda)$.

The above formulation is illustrated in Fig. 2, where Fig. 2(a) and (i) show a Laplacian speech decay rate distribution; Fig. 2(b) and (j) show a Gaussian room decay rate distribution, for one fast decay ($\lambda_h = -50$) and one slower decay ($\lambda_h = -10$) respectively, where λ_h is the true decay rate. Fig. 2(c), (d), (e), (f), (g) and Fig. 2(k), (l), (m), (n), (o) show the process $g(f_s(\tau), f_h(\lambda))/f_s(\tau)$ at the corresponding τ shown in Fig. 2(a) and (i), respectively. The *total* reverberant speech decay rate distribution is shown in Fig. 2(h) and (p). It can be seen that for the faster decay (smaller decay rate), the total distribution (depicted in Fig. 2(h)) shows a closer resemblance to that of the original Laplacian $f_s(\lambda)$, whereas for the slower decay (larger decay rate), the total distribution (depicted in Fig. 2(p)) shows a ‘skewed’ version of the original Laplacian $f_s(\lambda)$, as discussed in Section 4.

If the cumulative distribution of the room decay rate $F_h(\tau)$ up to the upper limit τ contains a significant portion of $f_h(\lambda)$, then the term $F_h(\tau)$ causes an impulse like contribution to the reverberant speech decay rate distribution. Note that the true decay rate λ_h is smaller for a faster decaying RIR. Therefore, at a given τ , the term $F_h(\tau)$ in $g(f_s(\tau), f_h(\lambda))$ will contain a larger portion of $f_h(\lambda)$ compared to a RIR with a slower decay. This will result in more impulse like contributions to the reverberant speech decay rate distribution as shown in Fig. 2(c), (d), (e), (f) and (g). If the terms $g(f_s(\tau), f_h(\lambda))/f_s(\tau)$ are equal to perfect impulses, then the sum of $g(f_s(\tau), f_h(\lambda))$ would be exactly that of the speech decay rate distribution, $f_s(\lambda)$. As the true room decay becomes slower, i.e. λ_h increases, less of the individual contribution $g(f_s(\tau), f_h(\lambda))/f_s(\tau)$ are impulse like. For example, Fig. 2(k) and (l) are not impulse like compared to the same corresponding $g(f_s(\tau), f_h(\lambda))/f_s(\tau)$ for Fig. 2(c) and (d). The latter contributions will cause the total distribution to be skewed. Hence, the relationship between the ‘skewness’ of $f_x(\lambda)$ and the true room decay rate, λ_h , can be used to predict the underlying reverberation time.

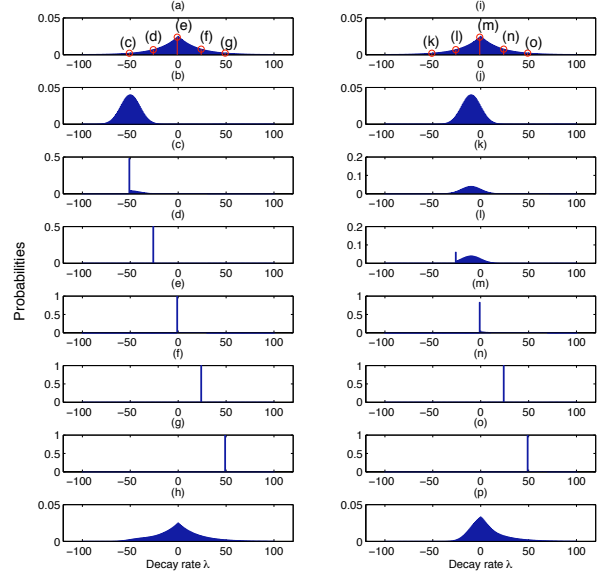


Fig. 2. Illustration of the decay rate distribution of the reverberant signal due to the max function of the room and speech decay rates.

4.2. Relation between ‘skewness’ and decay rate

There are many statistical measures that define the ‘skewness’ of a distribution, such as the third normalized central moment, skewness. We also employ the negative-side variance, denoted by σ_{x-}^2 . The negative-side variance is defined as the variance of a symmetrical distribution ($f_x^-(\lambda)$) with the same negative-side distribution of the original distribution ($f_x(\lambda)$),

$$f_x^-(\lambda) = \begin{cases} f_x(\lambda) & \text{for } \lambda \leq 0 \\ f_x(-\lambda) & \text{if } \lambda > 0. \end{cases} \quad (11)$$

In Fig. 3(a) and (b) the skewness and the negative-side variance are shown for different room decay rates λ_h and four different speakers. It should be noted that the skewness depends on both the positive- and negative-side variance. Since the positive side variance mostly depends on the distribution of the speech decay rate (see Fig. 2(h) and (p)) the skewness is more speaker dependent compared to the negative-side variance. Therefore, we propose to use the negative side variance to predict the room decay rate. A second order function was used map the observed σ_{x-}^2 , obtained from the reverberant speech decay rate distribution, to the estimated true room decay rate $\hat{\lambda}_h$ as:

$$\hat{\lambda}_h = \gamma_2(\sigma_{x-}^2)^2 + \gamma_1\sigma_{x-}^2 + \gamma_0 \quad (12)$$

The parameters ($\gamma_0, \gamma_1, \gamma_2$) of the mapping function were obtained by using Polack’s statistical reverberation model and two speech fragments consisting of one male and one female sentence. It should be noted that the parameters depend on the STFT and the decay rate LS fitting implementations.

5. EXPERIMENTAL METHOD AND RESULTS

The signals were first analysed using a standard STFT (Hamming window of 256 samples, FFT size of 512 samples, 75% overlap at 16 kHz). The decay rates that are used to generate the reverberant speech decay rate distribution were obtained by continuously applying a LS fit to 20 time frames (92 ms) for each frequency bin.

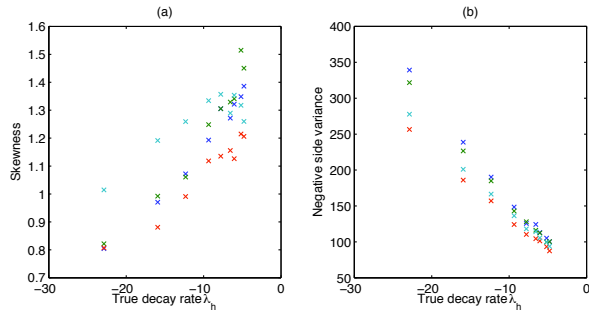


Fig. 3. Relationship between the skewness and the negative-side variance vs true decay rate.

The mapping function that relates the negative-side variance σ_x^2 to λ_h was validated using simulated and real recorded reverberant speech signals. The true RT is found from energy decay curve using Schroeder’s method [3] and used as a reference against which the estimates obtained from our method are compared. In Schroeder’s method, the energy decay curve of the RIR is first calculated by backward integration. Secondly, the decay rate is obtained by applying a linear LS fit in the range of -5 dB to -35 dB. Finally, the RT can be found using (2).

The simulated RIRs were generated using one realization of the Polack’s statistical reverberation model, and the image-method [12], for 10 different reverberation time between 0.1 to 1 s. The reverberant speech signals were then obtained by convolving the simulated RIRs with four speech signals which differ from the ones used for the calibration of the mapping parameters. In addition, two speech signals were recorded in two rooms with different acoustic properties with a reverberation time of 180 ms and 400 ms. The source-microphone distance was 4 m. In Fig. 4(a) the estimated reverberation times and the corresponding standard deviation across different speech signals is shown. In Fig. 4(b) the relative estimation error between the estimated and the true RT is shown. It can be seen in Fig. 4(b) that the proposed method can blindly estimate the RT an accuracy of about 5-15% for reverberant signals generated using Polack’s statistical reverberation model, and to about 15% for reverberant signals generated using the image-method. Furthermore, the results obtained using the recorded speech signals exhibits errors around 12% in RT estimation demonstrate the applicability of the method in a practical scenario.

It should be noted that the parameters of the mapping function were calibrated using Polack’s statistical reverberation model, which assumes that the reverberant field is diffuse. However, the recorded signals, and the simulated signals (generated using the image-method RIRs) do not exhibit a perfectly diffuse reverberant field. In general, this causes a steeper decay than the decay which is found using Schoeder’s method. Therefore, the RT is slightly underestimated when the RIRs are generated using the image-method (see Fig. 4(a)). The underestimation can also be seen for the results of two real rooms in Fig. 4(a). For some reverberation suppression techniques, e.g., [9], the underestimation of the RT causes only a slight decrease in the suppression performance, while overestimation would distort the speech signal.

6. CONCLUSION

In this paper a method was developed to estimate the reverberation time directly from the observed reverberant speech signal. The decay rates of the energy envelope of the signal are continuously estimated

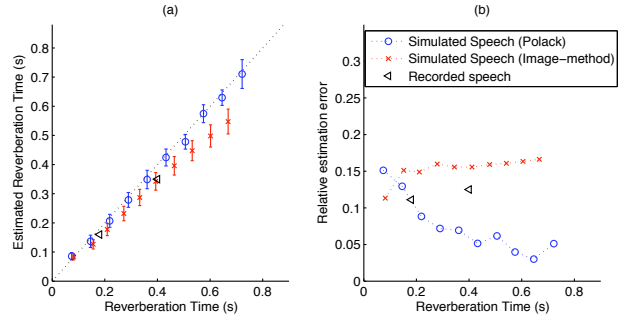


Fig. 4. Polack’s model: circle, Image-method: crosses, Real rooms: star. (a) Estimated RT against true RT, and its standard deviation across different speech fragments. (b) Relative estimation error for three different RIRs against the true RT.

in the STFT domain using a simple least squares fitting mechanism. The distributions of the room decay rate, the anechoic speech decay rate, and the reverberant speech decay rate were analysed. It was found that the negative-side variance of the reverberant speech decay rate distribution is a good measure for the true decay rate of the room. A second order mapping function was used to find the room decay rate given the negative-side variance of the reverberant speech decay rate distribution. The obtained decay rate is directly related to the reverberation time of the room. Experimental results have demonstrated the beneficial use of the developed method using simulated reverberant and real recorded speech signals.

7. REFERENCES

- [1] H. Kuttruff, *Room Acoustics*, 4th ed. Taylor & Francis, Oct. 2000.
- [2] ISO-3382, “Acoustics-measurement of the reverberation time of rooms with reference to other acoustical parameters,” *International Organization for Standardization, Gèneve*, 1997.
- [3] M. R. Schroeder, “A new method of measuring reverberation time,” *J. Acoust. Soc. Amer.*, vol. 37, pp. 409–412, 1965.
- [4] T. J. Cox, F. Li, and P. Darlington, “Extracting room reverberation time from speech using artificial neural networks,” *J. Audio Eng. Soc.*, vol. 49, no. 4, pp. 219–230, 2001.
- [5] K. Lebart, J. Boucher, and P. Denbigh, “A new method based on spectral subtraction for speech dereverberation,” *Acta Acustica*, vol. 87, pp. 359–366, 2001.
- [6] S. Vesa and A. Harma, “Automatic estimation of reverberation time from binaural signals,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, 2005, pp. 281 – 284.
- [7] R. Ratnam, D. L. Jones, B. C. Wheeler, J. William D. O’Brien, C. R. Lansing, and A. S. Feng, “Blind estimation of reverberation time,” *J. Acoust. Soc. Amer.*, vol. 114, no. 5, pp. 2877–2892, 2003.
- [8] J. D. Polack, “La transmission de l’énergie sonore dans les salles,” Ph.D. dissertation, Université du Maine, Le Mans, 1988.
- [9] E. A. P. Habets, “Single- and multi-microphone speech dereverberation using spectral enhancement,” Ph.D. dissertation, Technische Universiteit Eindhoven, 2007.
- [10] J. Y. C. Wen and P. A. Naylor, “An evaluation measure for reverberant speech using tail decay modeling,” in *Proc. European Signal Process. Conference*, 2006.
- [11] M. Unoki, M. Furukawa, K. Sakata, and M. Akagi, “A method based on the MTF concept for dereverberating the power envelope from the reverberant signal,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2003.
- [12] J. Allen and D. Berkley, “Image method for efficiently simulating small room acoustics,” *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.