EVALUATION OF A STOCHASTIC REVERBERATION MODEL BASED ON THE IMAGE SOURCE PRINCIPLE

Achille Aknin, Théophile Dupré, Roland Badeau

EVALUATION OF A STOCHASTIC REVERBERATION MODEL BASED ON THE IMAGE SOURCE PRINCIPLE

Achille Aknin
LTCI, Télécom Paris
Institut Polytechnique de Paris
Palaiseau, France
achille.aknin@telecom-paris.fr

Théophile Dupré
Aix-Marseille University
CNRS, PRISM
Marseille, France
dupre@prism.cnrs.fr

Roland Badeau
LTCI, Télécom Paris
Institut Polytechnique de Paris
Palaiseau, France
roland.badeau@telecom-paris.fr

ABSTRACT

Various audio signal processing applications, such as source separation and dereverberation, require an accurate mathematical modeling of the input audio data. In the literature, many works have focused on source signal modeling, while the reverberation model is often kept very simplistic.

This paper aims to investigate a stochastic room impulse response model presented in a previous article: this model is first adapted to discrete time, then we propose a parametric estimation algorithm, that we evaluate experimentally. Our results show that this algorithm is able to efficiently estimate the model parameters, in various experimental settings (various signal-to-noise ratios and absorption coefficients of the room walls).

1. INTRODUCTION

Audio signal processing algorithms often involve the modeling of room impulse responses. In the context of source separation or dereverberation, for example, the observed signal $x(t)$ is usually defined as a sum of convolution products of acoustic source signals $s_i(t)$ with the corresponding room impulse responses (RIR) $h_i(t)$, corrupted by additive noise $n(t)$ as in (1):

$$x(t) = \sum_i h_i * s_i(t) + n(t), \tag{1}$$

where i refers to the different acoustic sources. Hence, the respective mathematical models chosen for the source signals and the RIRs play an equally important role in the joint estimation of h_i and s_i.

The modeling of source signals has been the main focus in numerous papers, with approaches such as Non-negative Matrix Factorization (NMF) [1, 2], stochastic models [3], or methods that take advantage of specific characteristics of the problem, like harmonic/percussive separation [4].

As for the RIR, one standard model [5, 6] is a Gaussian process with independent samples and exponentially decreasing variance:

$$\begin{aligned} h(t) &= e^{-at} b(t) \\ b(t) &\sim \mathcal{N}(0, \sigma^2). \end{aligned} \tag{2}$$

We will show in Section 2 that this model is a good approximation in the late part of the reverberation but it does not permit to accurately represent the early reflections. However the early reflections, where the energy is mostly concentrated, are perceptually very important. Further attempts for modeling reverberation include the use of a spatial covariance matrix [7], complex Gaussian latent variables [8] or a more general Student's t model [3].

We present in this paper an algorithm that aims to estimate the parameters of a model which accurately represents both the early and late parts of reverberation, based on a previous work in [9, 10]. We intend to adapt this algorithm to various signal processing applications, such as source separation, noise reduction or dereverberation.

This paper is organized as follows: in Section 2 we define the stochastic model and discuss some of its properties, Section 3 exposes the parametric estimation algorithm and Section 4 reviews our experimental results. Finally, some conclusions and perspectives are drawn in Section 5.

2. REVERBERATION MODEL BASED ON A POISSON POINT PROCESS

The physical model we use is described in [9, 10]. We will summarize in this section the main contributions and highlight some interesting properties.

2.1. The image source principle

The model is based on the image source principle [11, 12], illustrated in Fig. 1. According to this principle, an indirect path from an acoustic source to a microphone can equivalently be described by a direct path from an image source to the microphone and conversely, where the image sources are at the positions obtained by iterative symmetrization of the original source with respect to the room walls.

A remarkable property of this principle is that, regardless of the room dimensions, the density of the image sources is uniform in the whole space: the number of image sources contained in a given disk, of radius sufficiently larger than the room dimensions, is approximately invariant under any translation of this disk. We will additionally consider that the positions of the image sources are random and uniformly distributed in the room. More precisely, for any given volume $V \subset \mathbb{R}^3$, we assume that the number of image sources in V follows a Poisson distribution (denoted $\mathcal{P}$) of parameter $\lambda|V|$:

$$N(V) \sim \mathcal{P}(\lambda|V|) \tag{3}$$

where $|.|$ denotes the Lebesgue measure.

This assumption leads to the use of a Poisson random measure with independent increments to describe the image sources, with

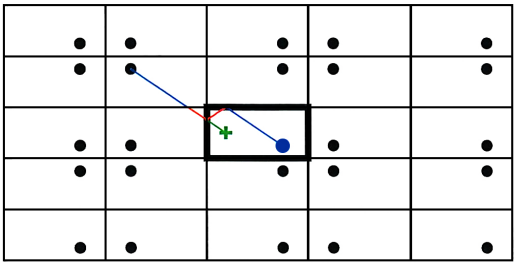


Figure 1: *Positions of microphone (plus sign), source (blue dot) and image sources (black dots) with the room walls represented as thick lines. A virtual straight trajectory from one image source to the microphone is drawn, along with the real trajectory in the original room [10].*

an infinitesimal volume $\mathrm{d}x_s$:

$$\begin{aligned} \mathrm{d}N(x_s) &\sim \mathcal{P}(\lambda \mathrm{d}x_s) \\ N(V) &= \int_V \mathrm{d}N(x_s) \end{aligned} \tag{4}$$

where we still have $N(V) \sim \mathcal{P}(\int_V \lambda \mathrm{d}x_s)) = \mathcal{P}(\lambda |V|)$.

2.2. Unified stochastic reverberation model

According to the image source principle, the RIR can be decomposed into a sum of direct sound waves from the image sources to the microphone:

$$h(t) = e^{-at}\tilde{h}(t) \text{ with } \tilde{h}(t) = \sum_s \tilde{h}_s(t)$$

where we sum over all the image sources s, and $a > 0$ is a constant exponential decay factor, assuming that the acoustic field is diffuse (isotropic) and that the absorption of the walls is independent of the frequency. We can further express the sound received from the image source s, $\tilde{h}_s(t)$, if we assume the source and the microphone to be omnidirectional:

$$\tilde{h}_s(t) = \frac{g\left(t - \frac{\|x - x_s\|_2}{c}\right)}{\|x - x_s\|_2}$$

where c is the speed of sound, $\|.\|_2$ denotes the Euclidean vector norm, x is the microphone position, x_s is the source position, $\frac{1}{\|x-x_s\|_2}$ is a consequence of the quadratic energy decay of spherical waves, and the causal filter g corresponds to other convolutive effects such as the inner response of the microphone, and must verify $\mathcal{F}_g(0) = \frac{\mathrm{d}\mathcal{F}_g}{\mathrm{d}f}(0) = 0$, where $\mathcal{F}_g(f)$ is the Fourier transform taken at frequency f, for technical reasons explained in [10]. Using the assumptions made in the last section, we can further develop:

$$\tilde{h}(t) = \int_{x_s \in \mathbb{R}^3} g\left(t - \frac{\|x - x_s\|_2}{c}\right) \frac{\mathrm{d}N(x_s)}{\|x - x_s\|_2}.$$

One final simplification can be made by noticing that, since the microphone and the source are omnidirectional, the integrated function only depends on the distance between the microphone and the image source $r = \|x - x_s\|_2 = ct'$, where t' is the time taken by a sound wave to travel from the image source to the microphone: $\forall t \in \mathbb{R}$,

$$\tilde{h}(t) = \int_{t' \in \mathbb{R}^+} g\left(t - t'\right) \frac{\mathrm{d}N'(t')}{ct'} \tag{5}$$

where $\mathbb{R}^+$ denotes the set of non-negative real numbers, and $\mathrm{d}N'(t') \sim \mathcal{P}(4\pi c^3 \lambda t'^2 \mathrm{d}t')$ is now a Poisson increment of quadratically increasing parameter w.r.t. time.

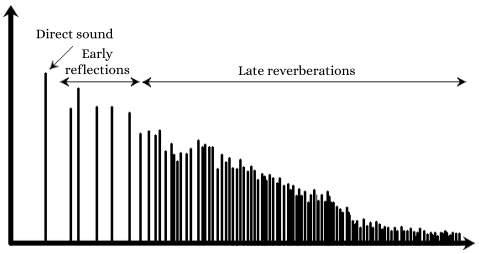


Figure 2: *Representation of an ideal RIR, with isolated peaks in the early reflections becoming denser in the late reverberation.*

2.3. Properties

The main difference between the exponentially decreasing Gaussian model in (2) and the Poisson point process model in (5) is the treatment of the early reflections. Fig. 2 shows that, in an ideal RIR, the early reflections correspond to isolated peaks which become denser with time, converging to a Gaussian process in the late part of reverberation.

While the exponentially decreasing Gaussian model is a good approximation in late reverberation, it does not accurately represent the early reflections. The Poisson point process model, on the other hand, is representative of both the early and late parts of the reverberation, thanks to the sparsity of the Poisson distribution.

2.4. Discrete-time model

From now on, we will consider the case of discrete-time signals, sampled at frequency f_s. Consequently the model (5) becomes: $\forall u \in [0, L_h - 1]$,

$$\begin{aligned} h(u) &= b(u) + w(u) \\ b(u) &= \sum_{v \in \mathbb{N}} g_d\left(u - v\right) \frac{e^{-a_d v} p(v)}{v} \end{aligned} \tag{6}$$

where:

- L_h is the length of the observed RIR h,
- $w(u) \sim \mathcal{N}(0, \sigma^2)$ is white Gaussian noise corresponding to the measurement error of h,
- $a_d = \frac{a}{f_s}$,
- $g_d(v) = e^{-a_d v} \frac{f_s}{c} g(\frac{v}{f_s})$,
- $p(v) \sim \mathcal{P}(\lambda_d v^2)$,
- $\lambda_d = 4\pi\lambda \frac{c^3}{f_s^3}$.

For readability purposes, we will respectively denote the parameters a, λ and g instead of a_d, λ_d and g_d and stop referring to their continuous equivalent in the rest of this paper.

3. ESTIMATING THE PARAMETERS

The model defined in (6) includes:

- the observed variable $h(u)$,
- the latent variable $b(u)$ [1],
- the parameters a, σ^2 and λ,
- the impulse response g that filters the time series $\pi(v) = \frac{p(v)}{v}$ scaled down by e^{-av}, that we will model as a causal autoregressive (AR) filter of parameters $(1, -\alpha_1, \dots, -\alpha_P)$,
- one hyper-parameter, the order P of the AR filter g.

To simplify the estimating process, we will consider $p(v)$ to be drawn from a Gaussian process of same variance, instead of a Poisson process. This invalidates some results from [10] but p still has the same physical interpretation. Thus we have:

$$\begin{aligned} p(v) &\sim \mathcal{N}(0, \lambda v^2) \\ \pi(v) &= \frac{p(v)}{v} \sim \mathcal{N}(0, \lambda) \end{aligned} \tag{7}$$

where $p(v)$ has an expected value of 0, instead of λv^2 as in Section 2.4, but this is compensated by the fact that we will no longer need to assume $\mathcal{F}_g(0) = \frac{\mathrm{d}\mathcal{F}_g}{\mathrm{d}f}(0) = 0$ as in Section 2.2. A consequence of this simplification is that the a posteriori distribution of b is also Gaussian:

$$b \mid h, \alpha, \lambda, a, \sigma^2 \sim \mathcal{N}(\mu, R). \tag{8}$$

Given these last conditions, an Expectation-Maximization (EM) algorithm [13] can be used to jointly estimate the parameters α, a, λ and σ^2 and the latent variable $b(u)$ given an observed RIR $h(u)$ in the maximum likelihood sense. The algorithm alternates two steps:

- expectation step (E-step): computing the a posteriori distribution of b given h (in this case we only need its mean vector μ and covariance matrix R), given the current estimates of the parameters,
- maximization step (M-step): maximizing (9) with respect to the parameters $\theta = (\alpha, \lambda, a, \sigma^2)$ given the current estimate of the a posteriori distribution of b.

More specifically, the a posteriori expectation of the log-probability density function (PDF) of the joint distribution of observed and latent random variables is:

$$\begin{aligned} \mathcal{Q} =& \mathbb{E}_{\mathbb{P}(b|h,\theta)} \left[\ln \mathbb{P}(h, b \mid \theta)\right] \\ =& \mathbb{E}_{\mathbb{P}(b|h,\theta)} \left[\ln \mathbb{P}(b \mid \theta)\right] + \mathbb{E}_{\mathbb{P}(b|h,\theta)} \left[\ln \mathbb{P}(h \mid b, \theta)\right] \\ =& -\frac{L_h}{2} \ln(2\pi\lambda) + \frac{L_h(L_h-1)}{2} a \\ & - \frac{1}{2\lambda} \left(\|EA\mu\|_2^2 + \mathrm{Tr}(E^2 A R A^T) \right) \\ & - \frac{L_h}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left(\|h - \mu\|_2^2 + \mathrm{Tr}(R) \right) \end{aligned} \tag{9}$$

with:

- M^T the transpose of matrix M,
- $\mathrm{Tr}(M)$ the trace of matrix M,
- E the diagonal matrix of coefficients $\{e^{au}\}_{u=0}^{L_h-1}$,
- A the $L_h \times L_h$ Toeplitz matrix implementing the inverse filter of the AR filter g, which has finite impulse response:

$$A = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ -\alpha_1 & 1 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ -\alpha_P & & \ddots & \ddots & \ddots & \vdots \\ & \ddots & & \ddots & 1 & 0 \\ 0 & & -\alpha_P & \dots & -\alpha_1 & 1 \end{pmatrix}.$$

3.1. Expectation

The expectation step corresponds to the exact updates:

$$\begin{aligned} R &= \lambda\sigma^2(\lambda I + \sigma^2 A^T E^2 A)^{-1} \\ \mu &= \frac{Rh}{\sigma^2} \end{aligned}$$

with I the identity matrix. Since the matrix inversion in the update of R is computationally expensive, using a Kalman filter with a Rauch-Tung-Striebel (RTS) smoother as described in [14] and [15] is preferable.

In this case, considering at each time step $u \geq 0$ the state variable $B(u) = \left[b(u), b(u-1), \dots, b(u-P+1), b(u-P)\right]^T \sim \mathcal{N}(\mu_{u|L_h-1}, R_{u|L_h-1})$ (where $\mu_{u|v}$ and $R_{u|v}$ denote the mean vector and covariance matrix of $B(u)$ given all observations h from times 0 up to v), the following two equations hold for any $u > 0$:

$$\begin{aligned} B(u) &= FB(u-1) + \Pi(u) \\ h(u) &= CB(u) + w(u) \end{aligned}$$

with:

- $\Pi(u) \sim \mathcal{N}(0, Q_u)$,
- $Q_u(0,0) = \lambda e^{-2au}$ and $Q_u(i,j) = 0$ if $i \neq 0$ or $j \neq 0$,
- $F = \begin{pmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_p & 0 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix}$,
- $C = [1, 0, \dots, 0]$.

The expectation step is then as described in Algorithm 1.

3.2. Maximization

As for the maximization step, the following updates directly maximize the a posteriori expectation of the joint log-PDF $\mathcal{Q}$ in (9).

[1] Note that we could equivalently define $w(u)$ or $p(v)$ as the latent variable.

Algorithm 1: Expectation step with Kalman filter and RTS smoother

Result: $\mu_{u|L_h-1}$ and $R_{u|L_h-1}$ for all $u \geq 0$
$\mu_{-1|-1} = [0, \ldots, 0]$;
$R_{-1|-1} = \mathbf{0}$ the zero matrix;
for $u \leftarrow 0$ **to** $L_h - 1$ **do**
  $\mu_{u|u-1} = F\mu_{u-1|u-1}$ (a priori state estimate);
  $R_{u|u-1} = FR_{u-1|u-1}F^T + Q_u$ (a priori covariance estimate);
  $\tilde{y}_u = h(u) - C\mu_{u|u-1}$ (pre-fit residual);
  $S_u = CR_{u|u-1}C^T + \sigma^2$ (pre-fit residual covariance);
  $K_u = R_{u|u-1}C^T/S_u$ (optimal Kalman gain);
  $\mu_{u|u} = \mu_{u|u-1} + K_u\tilde{y}_u$ (a posteriori state estimate);
  $R_{u|u} = (I - K_uC)R_{u|u-1}$ (a posteriori covariance estimate);
end
for $u \leftarrow L_h - 1$ **to** 1 **do**
  $J_{u-1} = R_{u-1|u-1}F^T R^{-1}_{u|u-1}$ (correction matrix);
  $\mu_{u-1|L_h-1} = \mu_{u-1|u-1} + J_{u-1}(\mu_{u|L_h-1} - \mu_{u|u-1})$ (smoothed state estimate);
  $R_{u-1|L_h-1} = R_{u-1|u-1} + J_{u-1}(R_{u|L_h-1} - R_{u|u-1})J^T_{u-1}$ (smoothed covariance estimate);
end

Filter g: The optimal filter parameters are the unique solution of the linear system: $\forall 0 < p \leq P$,

$$\sum_{q=1}^{P} \alpha_q \sum_{u=0}^{L_h-1} (\mu(u-q)\mu(u-p) + R(u-q, u-p))e^{2au} = \sum_{u=0}^{L_h-1} (\mu(u)\mu(u-p) + R(u, u-p))e^{2au}. \tag{10}$$

Image sources density parameter λ: The optimal parameter λ is:

$$\lambda = \frac{1}{L_h}(\|EA\mu\|_2^2 + \text{Tr}(E^2ARA^T)). \tag{11}$$

Absorption parameter a: Substituting the value of λ in the expression of $\mathcal{Q}$ in (9) and canceling the partial derivative with respect to a, we find:

$$\frac{L_h - 1}{2}\left[\|EA\mu\|_2^2 + \text{Tr}(E^2ARA^T)\right] = \sum_{u=0}^{L_h-1} u\left[(EA\mu)(u)^2 + (E^2ARA^T)(u,u)\right]. \tag{12}$$

This equation has no closed-form solution, but the solution is unique and we can use a dichotomy method to find the optimal value of a.

White noise parameter σ^2:
The update of the white noise parameter is:

$$\sigma^2 = \frac{1}{L_h}\left(\|h - \mu\|_2^2 + \text{Tr}(R)\right). \tag{13}$$

3.3. Initialization of the parameters

An adequate initialization of the parameters can make the EM algorithm converge much faster. This is why we initialize the parameters as follows:

Filter g:
Filter g has an effect on the spectral shape of h, so that we can approximate the squared magnitude of the frequency response of g by taking the average power of each sub-band signal in the short-term Fourier transform (STFT) of h. The inverse discrete Fourier transform (DFT) of this squared magnitude of the frequency response gives a biased estimate of the autocovariance function, and the corresponding AR coefficients α are found by solving Yule-Walker's equations [16].

Absorption parameter a:
Since the energy of h is exponentially decreasing, we perform linear regression on $\ln h^2(u)$ in order to estimate its slope $2a$.

Image sources density parameter λ:
Given that the approximate number of impulses in the early part of the RIR between times 0 and N is $\lambda\frac{N^3}{3}$, we consider the first sample N_{first} where h reaches a certain threshold (for example half the maximum amplitude of h), which is expected to correspond to the first impulse (due to the direct path from the source to the sensor). We then approximate $\lambda = \frac{3}{N^3_{first}}$. Although inaccurate, this approximation is sufficient for initializating the EM algorithm.

White noise parameter σ^2:
The white noise parameter σ^2 is estimated as the average power of the first values of h, before the first impulse at time N_{first}.

3.4. Implementation

The algorithm described in this section was implemented in Python. Code and documentation can be found in the following GitHub repository: https://github.com/Aknin/Gaussian-ISP-model.

4. EXPERIMENTAL RESULTS

4.1. Room impulse response dataset

In order to evaluate the performance of the EM algorithm, we used synthetic RIRs generated by Roomsimove [17]. Roomsimove is a MATLAB toolbox that simulates a parallelepipedic room, and allows us to specify the dimensions of the room, the absorption parameter of the walls, the filter g, and the position of the source and microphone. We manually added the noise $w(u)$ to account for the measurement error. Although the model described in [9, 10] applies to any room geometry and our algorithm is expected to work regardless of the shape of the room, Roomsimove is limited to the simulation of parallelepipedic rooms.

In all the experiments conducted, we used the following parameters in Roomsimove:

- g_{true} is a combination of a low-pass filter with cut-off frequency 20 Hz, that is implemented as a recursive filter of order (2,2) (default settings of Roomsimove), and a Finite Impulse Response (FIR) filter approaching the frequency response of an Audio-Technica ATM650 microphone[2],

[2] Based on http://recordinghacks.com/microphones/Audio-Technica/ATM650.

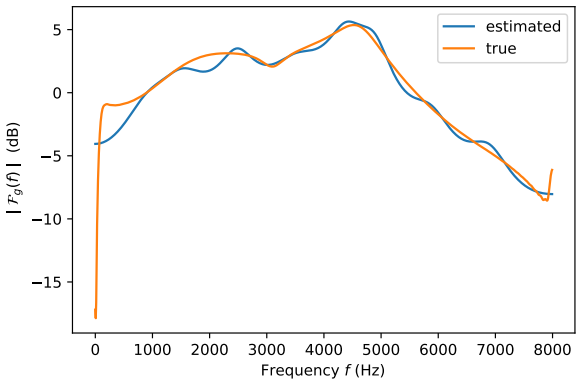


Figure 3: Frequency response of both the estimated and true filters g in one of the experiments, where the absorption coefficient is 0.7 and $\sigma^2 = 10^{-4}$ (the corresponding SNR is 15 dB).

- the room is of size $2 \times 3 \times 4$ (in meters),
- the sampling frequency is $f_s = 16000$ Hz,
- the sources and microphones in the room are omnidirectional.

We also set the hyper-parameter $P = 20$ (order of the AR filter).

Note that, since to the best of our knowledge the algorithm presented in this paper is the first one that estimates the model presented in [10], it is not possible to compare its performance with any other method in the literature. Instead, we will use the true parameters of Roomsimove and compare them to the estimated parameters. For instance, the comparison of the true and estimated σ^2 is straightforward, since the noise is manually added.

We can also compare the estimate g_{est} and the ground truth g_{true} by computing the average relative error between their respective DFTs:

$$D_r(g_{est}, g_{true}) = \frac{1}{N_{fft}-1} \sum_{k=1}^{N_{fft}-1} \frac{\left|\left|\hat{G}_{est}(k)\right| - \left|\hat{G}_{true}(k)\right|\right|}{\left|\hat{G}_{true}(k)\right|}$$

where N_{fft} is the number of frequency bins k in the DFT $\hat{G}(k)$. Note that the first frequency bin $k = 0$ is ignored because, as an AR filter, g_{est} is not expected to verify the property satisfied by g_{true}: $\mathcal{F}_{g_{true}}(0) = \frac{\mathrm{d}\mathcal{F}_{g_{true}}}{\mathrm{d}f}(0) = 0$. Fig. 3 shows an example of the frequency response of the estimated g_{est} along with the frequency response of the true filter g_{true} where $D_r(g_{est}, g_{true}) = 0.151$.

Parameter a is related to the reverberation time T_{60} of the room ($T_{60,est} = \frac{3\ln(10)}{a f_s}$) which can be computed with Roomsimove parameters using Eyring's formula [18]:

$$T_{60,true} = -0.1611 \frac{|V|}{S \ln(1-A)}$$

where $|V|$ is the volume of the room, S is the total surface of the walls in the room and A is the absorption of the walls. We can also estimate $T_{60,baseline}$, that will serve as a baseline, by interpolating the logarithm of the Energy Decay Curve (EDC) with a linear function of coefficient $-T_{60,baseline}$[3] as described in [19].

As for parameter λ, it is theoretically related to the volume of the room: $\lambda = \frac{1}{|V|}$, and it is directly related to the energy of h. But since the observed RIR h is normalized in practice, we cannot compare λ to any ground truth parameter.

4.2. Complexity of the algorithm

While the exact expectation step, involving the inversion of an $L_h \times L_h$ matrix, is quite computationally expensive ($\mathcal{O}(L_h^3)$), the complexity of the Kalman filtering and RTS smoothing algorithm is only $\mathcal{O}(L_h P^3)$.

Note that the maximization step is even less computationally expensive. Indeed, the most expensive stage in this M-step is the computation of the diagonal entries of matrix ARA^T in (11) and (12). Knowing that a multiplication Ax (or $x^T A^T$) with x a vector of length L_h is actually a convolution of x with a finite impulse response of length $P+1$, the computation of these diagonal entries amounts to a complexity of $\mathcal{O}(L_h P^2)$.

In practice, on an Intel Xeon Gold 6154 CPU at 3.00 GHz, the execution time for 500 iterations of the EM algorithm is 150 seconds, in the case of short RIRs of length $L_h = 750$ (i.e. 47 ms), or up to 600 seconds in the case of longer RIRs of length $L_h = 2500$ (i.e. 156 ms). It is important to observe that only 150 iterations are sufficient to converge to accurate parameter estimates when $L_h = 750$, but more iterations are needed when the reverberation in longer.

4.3. Influence of the Signal to Noise Ratio on the estimation

In this experiment, we made the white noise parameter σ^2_{true} vary in order to investigate the influence of the signal-to-noise ratio (SNR) on the estimation performance, while fixing the absorption parameter to 0.7.

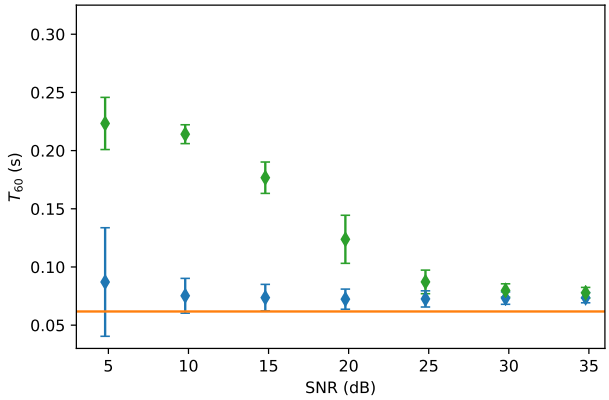
Fig. 4 compares the true and estimated parameters for several SNRs between 5 and 35. Fig. 4c, that compares σ^2_{true} and σ^2_{est}, displays the quotient of the estimate over the true parameter, in order to improve the readability (σ^2 takes values between 10^{-6} and 10^{-3}). The blue dot represents the mean value and the blue segment is bounded by this mean value plus or minus the standard deviation, where the mean and standard deviations are computed over 100 different experiments for each SNR value. Fig. 4a also shows the baseline estimation $T_{60,baseline}$ using green dots and green segments in a similar way.

We can draw several conclusions from these figures. First of all, the estimation of the T_{60} seems to be better than the baseline estimation, although slightly biased at any SNR value. Having a high SNR gives more consistent results, but the estimation remains biased.

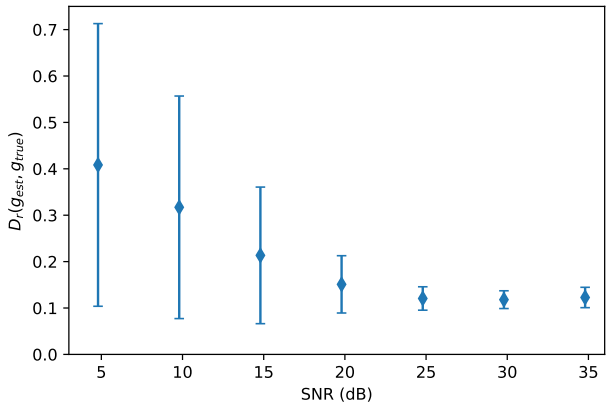
On the other hand, a high SNR leads to better estimates of g, by improving both the consistency and the mean results.

As for the estimation of σ^2, it fits well the true values regardless of the SNR, except for a higher standard deviation when the SNR is too low.

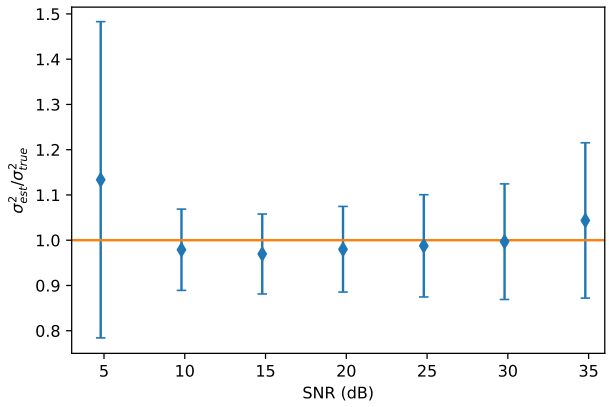
[3] Note that being able to estimate the T_{60} does not allow us to compare this algorithm to other T_{60} estimators available in the literature, since they are usually designed to work with speech or music signals, instead of RIRs, and generally fail when applied to RIRs.



(a) Comparison of $T_{60,est}$ (blue), $T_{60,baseline}$ (green) and $T_{60,true}$ (orange)

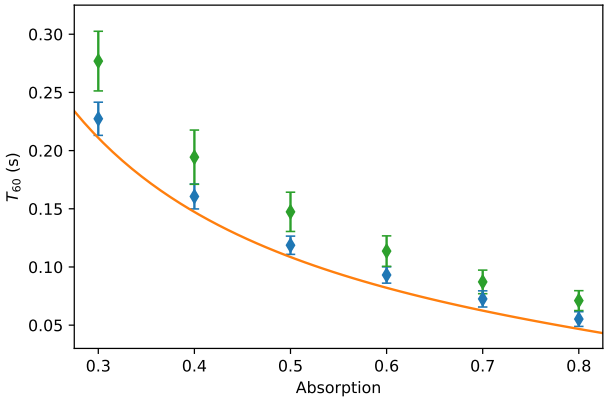


(b) $D_r(g_{est}, g_{true})$

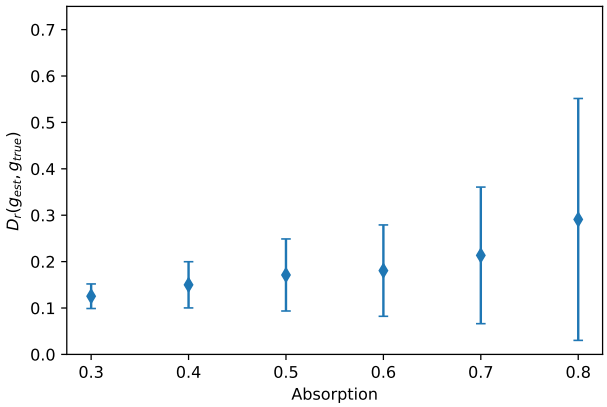


(c) Comparison of $\frac{\sigma^2_{est}}{\sigma^2_{true}}$ and 1

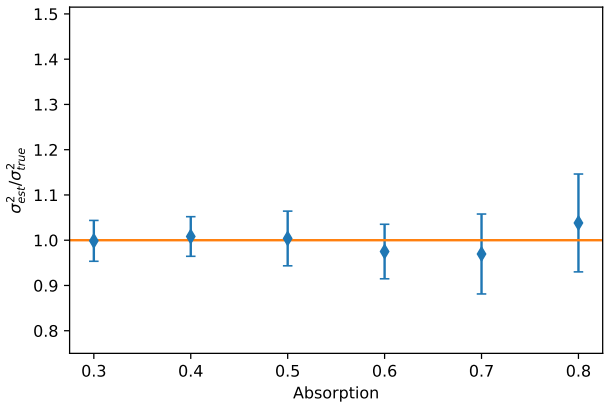
Figure 4: *Comparison of the mean and standard deviation of the estimation over* 100 *different experiments (blue) to the true parameters (orange) as well as the baseline estimation of the* T_{60} *(green) for different SNRs.*



(a) Comparison of $T_{60,est}$ (blue), $T_{60,baseline}$ (green) and $T_{60,true}$ (orange)



(b) $D_r(g_{est}, g_{true})$



(c) Comparison of $\frac{\sigma^2_{est}}{\sigma^2_{true}}$ and 1

Figure 5: *Comparison of the mean and standard deviation of the estimation over* 100 *different experiments (blue) to the true parameters (orange) as well as the baseline estimation of the* T_{60} *(green) for different absorption parameters.*

4.4. Influence of the absorption of the walls on the estimation

In a second experiment, we explored the influence of the absorption parameter of the room walls on the quality of the estimation. Fig. 5 shows the results obtained when the absorption ranges from 0.3 to 0.8 while σ^2_{true} is set to 10^{-5}, corresponding to an SNR of 15 dB.

In Fig. 5a we notice that the estimation of the reverberation time is closely matching the curve of $T_{60,true}$, while still being slightly biased. As in the first experiment, our model estimates the reverberation time better than the baseline.

On the other hand, a low absorption leads to better estimates of g, by improving both the consistency and the mean results.

As for the estimation of σ^2, the estimation appears to be good at any absorption value. This experiment thus confirms that the EM algorithm is robust to various absorption levels.

5. CONCLUSION

This paper investigated the application of a stochastic reverberation model and the estimation of its parameters given an observed RIR. We presented the implementation of the estimating algorithm and highlighted some experimental results.

In the future, this algorithm could be adapted to more specific tasks, such as estimating the T_{60} in various settings (since it is directly related to one of the parameters of the model), or also more complex signal processing applications, like source separation and dereverberation.

We now plan to further explore several directions:

- relaxing the assumption that the exponential decay is isotropic and not frequency dependent, leading to a more realistic reverberation model [20],
- adapting the EM algorithm to consider non-omnidirectional sources and microphones,
- implementing a similar algorithm with non-Gaussian processes, into order to better account for the sparsity of early reflections,
- estimating the model parameters from real audio signals (speech, music) instead of the RIR,
- implementing an audio source separation algorithm using this reverberation model in order to further evaluate the benefits of using an accurate representation of the RIR.

6. REFERENCES

[1] Alexey Ozerov and Cédric Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[2] Hiroshi Sawada, Hirokazu Kameoka, Shoko Araki, and Naonori Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[3] Simon Leglaive, Roland Badeau, and Gaël Richard, "Student's t source and mixing models for multichannel audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1154–1168, 2018.

[4] Derry Fitzgerald, "Harmonic/percussive separation using median filtering," *International Conference on Digital Audio Effects (DAFX10), Graz, Austria*, vol. 13, pp. 217–220, 2010.

[5] Manfred R. Schroeder, "Frequency-correlation functions of frequency responses in rooms," *The Journal of the Acoustical Society of America*, vol. 34, no. 12, pp. 1819–1823, 1962.

[6] James A. Moorer, "About this reverberation business," *Computer Music Journal*, vol. 3, no. 2, pp. 13–28, 1979.

[7] Ngoc QK Duong, Emmanuel Vincent, and Rémi Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[8] Laurent Girin and Roland Badeau, "On the use of latent mixing filters in audio source separation," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2017, pp. 225–235.

[9] Roland Badeau, "Unified stochastic reverberation modeling," in *Proc. of 26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, Sept. 2018.

[10] Roland Badeau, "Common mathematical framework for stochastic reverberation models," *The Journal of the Acoustical Society of America*, 2019, Special issue on room acoustics modeling and auralization.

[11] Heinrich Kuttruff, *Room acoustics*, CRC Press, 2016.

[12] Jont B. Allen and David A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[14] Greg Welch, Gary Bishop, et al., "An introduction to the Kalman filter," Tech. Rep. TR 95-041, University of North Carolina, Department of Computer Science, 1995.

[15] Herbert E. Rauch, F. Tung, and Charlotte T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA journal*, vol. 3, no. 8, pp. 1445–1450, 1965.

[16] Petre Stoica and Randolph Moses, *Spectral Analysis of Signals*, Prentice Hall, 2005.

[17] Emmanuel Vincent and Douglas R. Campbell, "Roomsimove toolbox," GNU Public License https://members.loria.fr/EVincent/software-and-data/, 2008.

[18] G Millington, "A modified formula for reverberation," *The Journal of the Acoustical society of America*, vol. 4, no. 1A, pp. 69–82, 1932.

[19] Heinrich W Löllmann and Peter Vary, "Estimation of the frequency dependent reverberation time by means of warped filter-banks," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 309–312.

[20] Roland Badeau, "Stochastic reverberation model for uniform and non-diffuse acoustic fields," Tech. Rep. 2019D003, Télécom ParisTech, Paris, France, 2019.