Blind estimation of room acoustic parameters using kernel regression

Arthur Belhomme\textsuperscript{1,2}, Yves Grenier\textsuperscript{1}, Roland Badeau\textsuperscript{1} and Eric Humbert\textsuperscript{2}

\textsuperscript{1}Institut Mines-Télém\textsuperscript{o}com; Télécom ParisTech; CNRS LTCI, Paris, 75013, France
\textsuperscript{2}Invoxia, Issy-les-Moulineaux, 92130, France

Correspondence should be addressed to Arthur Belhomme (arthur.belhomme@telecom-paristech.fr)

ABSTRACT
Room acoustic parameters are key information for dereverberation or speech recognition. Usually, when one needs to assess the level of reverberation, only the reverberation time $RT_{60}$ or a direct to reverberant sounds index $D_\tau$ is estimated. Yet, methods which blindly estimate the reverberation time from reverberant recorded speech do not always differentiate the $RT_{60}$ from the $D_\tau$ to evaluate the level of reverberation. That is why we propose a method to jointly blindly estimate these parameters, from the signal energy decay rate distribution, by means of kernel regression. Evaluation is carried out with real and simulated room impulse responses to generate noise-free reverberant speech signals. The results show this new method outperforms baseline approaches in our evaluation.

1. INTRODUCTION

When a sound is emitted in an enclosed space, the microphone does not only capture the output of the source: all the paths the sound may follow, from the source to the microphone, are added to the direct one and produce reverberation. These paths depend on the different reflections of the acoustic wave on the walls and surfaces it meets, that can be viewed by means of the room impulse response (RIR). This transfer function of the room would be the record of an impulse played from a spherical source and depends both on the room characteristics and the source-microphone distance.

Many parameters can be extracted from the RIR, however the reverberation time ($RT_{60}$) is the most common to describe the reverberation of the room. That is why one can find a large amount of methods in the literature to compute it from a measured RIR ([1], [2]) or estimate it blindly from a recording in a room. Among them, a major stream uses a maximum-likelihood procedure to estimate the time-constant of the signal energy decay, directly related to the $RT_{60}$ ([3], [4], [5], [6], [7]). Other methods use a machine learning approach by training deep neural networks in order to map the spectrum envelopes of reverberant signals to the corresponding room acoustic parameters ([8], [9], [10]). Finally, a bench of methods use the statistical distribution of the decay rate of the energy envelope, in order to link some statistical moments with the reverberation time ([11], [12], [13], [14]); this stream is the most similar to our approach.

However the reverberation time is specific to the characteristics of the room, its dimensions and material, but independent of the source-receiver distance. Yet this distance has a great impact on the amount of reverberation [15], characterized by the ratio of direct to reverberant sounds. Even if the direct-to-reverberant ratio (DRR) is the most straightforward index to assess the level of reverberation [16] many variants of this measure exist, including the clarity and definition indexes, well-suited for syllable and music intelligibility [17]. Thus, for a given room (i.e. for a given $RT_{60}$) one can obtain a certain range of clarity or definition index, according to the distance between the source and the microphone. Thereby, methods such as [12] and [11] which aim to link the $RT_{60}$ to a single feature of the energy decay rate distribution, through a fixed $n^{th}$-order polynomial, will not distinguish a low $RT_{60}$ and a high source-microphone distance from a high $RT_{60}$ and a short source-microphone distance. To fully characterize reverberation, one thus needs to estimate the $RT_{60}$ and an index related to the DRR.

On that account, we propose a method to jointly estimate the reverberation time and a clarity (or definition) index, from the energy decay rate distribution of a re-
verberant signal. We use kernel regression, specifically the Nadaraya-Watson estimator [18], on a database of energy rate distributions obtained from a wide range of synthetic and real RIRs. The paper is organised as follows: Section 2 presents the room acoustic parameters to be estimated and the used features. In Section 3 we derive the estimator and present our method. The evaluation database is detailed in Section 4 and the results are compared to state of the art methods ([19],[15]) in Section 5. Finally, in Section 6 some conclusions are drawn.

2. FEATURE EXTRACTION AND ROOM ACOUSTIC PARAMETERS

2.1. Room acoustic parameters

As claimed in Section 1, our goal is to jointly estimate the reverberation time and an index of the reverberation level. We define in this section the different expressions of these parameters.

Reverberation time

Defined as the time interval to measure an energy decay of 60 dB, the RT₆₀ can be computed with the Schroeder backward integration method [2] via the Energy Decay Curve (EDC):

$$\text{EDC}(n) = \sum_{k=n}^{N_h} h(k)^2, \forall n \geq 0$$

where $h$ is the room impulse response of length $N_h$. Then, the RT₆₀ is the required time for the EDC to decrease by 60 dB.

Level of reverberation

Two close indexes to assess the balance between the direct and reverberant sounds are the clarity index, defined in [16]

$$C_\tau = 10 \log_{10} \left( \frac{\sum_{n=0}^{N_r} h^2(n)}{\sum_{n=N_r+1}^{N_h} h^2(n)} \right) \text{ dB},$$

and the definition index, defined in [16]

$$D_\tau = 10 \log_{10} \left( \frac{\sum_{n=0}^{N_r} h^2(n)}{\sum_{n=0}^{N_h} h^2(n)} \right) \text{ dB},$$

where $N_r$ is the number of samples corresponding to the time $\tau$ ranging from 0.1 ms to 1 s, and $h(n)$ is the room impulse response. Two widely adopted values of $\tau$ are 50 ms or 80 ms, since they correspond to the duration splitting the (useful) early reflections to the (disturbing) late reverberation [20]. As explained in [17], the D₅₀ is an objective criterion to measure the speech intelligibility, while de C₅₀ is more designed for music transparency. However Parada et al. showed in [15] that the C₅₀ is more correlated with the Perceptual Evaluation of Speech Quality (PESQ) and the phoneme accuracy rate than the D₅₀. Even if the performance of our method is quite the same for the C₅₀ and D₅₀, we focus on the C₅₀ in order to later compare our results, as there is no other method for D₅₀ estimation to our knowledge.

2.2. Feature extraction

2.2.1. Decay rate distribution

The statistical reverberation model developed by Polack in [21] is often used to describe RIRs in a diffuse field. The RIR is represented as one realization of a non-stationary stochastic process, a Gaussian white noise of variance $\sigma^2$ damped by a decreasing exponential envelope, linked to a room parameter $\delta$:

$$h(n) = b(n)e^{-\delta \frac{n}{\text{RT}_60}}$$

with $b(n) \sim N(0, \sigma^2)$, $\delta = \frac{3 \ln 10}{\text{RT}_60}$ and $f_s$ the sampling rate. This model has been generalized by Habets to the entire RIR in [20], using a different noise (Gaussian with another variance) for the direct path. The energy envelope of the RIR can be expressed as:

$$e(n) = \mathbb{E}[h(n)^2] = \sigma^2 e^{-2\delta \frac{n}{\text{RT}_60}} = \sigma^2 e^{\lambda n}$$

with $\mathbb{E}[]$ the expectation operator and $\lambda_n = -2\delta$ the energy decay of the room.

However, we aim to analyze a reverberant speech, which can be viewed as the convolution of a RIR and an anechoic speech. Wen et al. derived in [11] an expression of the energy envelope of a reverberant signal $d_s(n)$, after a speech endpoint according to the energy decay rates of the anechoic signal and the room, $\lambda_s$ and $\lambda_h$ respectively:

$$d_s(n) = \begin{cases} 
\frac{(e^{\lambda_h n} - e^{\lambda_s n})}{ne^{\lambda_s n}}/(\lambda_h - \lambda_s) & \text{if } \lambda_h \neq \lambda_s \\
ne^{\lambda_s n} & \text{if } \lambda_h = \lambda_s 
\end{cases}$$

As the sum of two exponential terms will be dominated by the exponential term with the largest value $\lambda$, the en-
2.2.2 Characteristic function

The methods developed in [11] and [12] use the negative side variance of the distribution (i.e. the variance where the negative values of \( \rho \) have been symmetrized) to estimate the RT\(_{60}\) with a polynomial model built on a training dataset. In [13], the authors prefer to use the temporal differential of cepstral coefficients and collect the variance, skewness, kurtosis, median absolute deviation of the obtained distribution as input features for support vector regression.

Therefore, the relevance of some statistical moments of the reverberant energy rate distribution, for room acoustic parameters estimation, is clear. However the choice of the statistical moments to be used is unclear, nor shared between the authors. This is why we choose to use the characteristic function of the energy decay rate distribution, as it conveys the global information of the distribution, and indirectly contains all the statistical ordinary moments. As presented in [22], the characteristic function of a real random variable \( X \) is defined as:

\[
\phi_X(f) = \int e^{ifx} dF_X(x) = \mathbb{E}[e^{ifX}]
\]

where \( F_X(x) \) is the cumulative distribution of \( X \) and \( f \) is the angular frequency. As the characteristic function behaves simply under shift, scale changes and summation of independent variables, it is a convenient tool for parameters estimation [22]. Moreover, the different statistical ordinary moments, related to the room acoustic parameters, can be extracted from this function. If a random variable has ordinary moments up to the \( k^{th} \) order, the characteristic function has a \( k^{th} \) derivative at the zero frequency and:

\[
\mathbb{E}[X^k] = (-i)^k \phi_X^{(k)}(0)
\]

where \( \phi_X^{(k)} \) is the \( k^{th} \) derivative of \( \phi_X \). This is why we decide to represent the distribution of the energy decay rate of reverberant speech by its characteristic function. It will constitute the observation feature in a kernel regression approach.

3. BLIND ESTIMATION OF ACOUSTIC ROOM PARAMETERS

3.1 The Nadaraya-Watson estimator

Consider a random input vector \( X \) and a random output vector \( Y \), with joint probability density function (JPDF) \( P_{X,Y}(x,y) \). We seek a function \( f \) which best predicts \( Y \) given \( X \), by minimizing the expected prediction error (EPE) \( \mathbb{E} \left[ (Y - f(X))^2 \right] \). The solution is given by the regression function [18]:

\[
f(x) = \mathbb{E}[Y|X = x] = \int y P_{X,Y}(y|x) dy.
\]

A way to estimate this regression function is to use a kernel to approximate the JPDF. As defined in [23], a kernel \( K \) is a similarity measure of the form

\[
K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \\
(x,x') \mapsto <\phi(x), \phi(x')>
\]

where \( \phi \) is a mapping function from the input space \( \mathcal{X} \) to a dot product space \( \mathcal{V} \):

\[
\phi : \mathcal{X} \rightarrow \mathcal{V} \\
x \mapsto \phi(x)
\]

In regression applications, kernels are parameterized by a constant \( \lambda \) that dictates the width of the neighborhood to be considered, and are denoted \( K_\lambda \). With \( N \) realizations \((x_i,y_i)\) of the random variables \((X,Y)\), and a kernel \( K_\lambda \), one can estimate the JPDF \( \hat{P}_{X,Y}(x,y) \) [18] as:

\[
\hat{P}_{X,Y}(x,y) = \frac{1}{N} \sum_{i=1}^{N} K_\lambda(x,x_i) K_\lambda(y,y_i).
\]
Let us recall (11), noticing $P_{X,Y}(y|x) = P_{X,Y}(x,y)/P_X(x)$ and $P_X(x) = \int P_{X,Y}(x,y)dy$, we obtain:

$$f(x) = \frac{\int y P_{X,Y}(x,y)dy}{\int P_{X,Y}(x,y)dy}. \quad \text{(13)}$$

If we use the estimator (12) we can approximate the regression function by:

$$\hat{f}(x) = \frac{\sum_{i=1}^{N} K_\lambda(x,x_i)K_\lambda(y,y_i)dy}{\sum_{i=1}^{N} K_\lambda(x,x_i)\int K_\lambda(y,y_i)dy} = \frac{\sum_{i=1}^{N} K_\lambda(x,x_i)\int y K_\lambda(y,y_i)dy}{\sum_{i=1}^{N} K_\lambda(x,x_i)\int K_\lambda(y,y_i)dy}. \quad \text{(14)}$$

With the properties of kernel functions proposed in [23], $\int y K_\lambda(y,y_i)dy = y_i$ and $\int K_\lambda(y,y_i)dy = 1$, which leads to the Nadaraya-Watson estimator:

$$\hat{f}(x) = \frac{\sum_{i=1}^{N} y_i K_\lambda(x,x_i)}{\sum_{i=1}^{N} K_\lambda(x,x_i)}. \quad \text{(15)}$$

### 3.2 Features and output vectors

Thus, our method needs a dataset of $N$ pairs $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}^2$ where each $x_i$ is the characteristic function $\phi_\rho(f)$ of energy decay rate distribution of a reverberant speech, and $y_i$ is the corresponding room acoustic parameters: $(RT_{60},C_{50})$ or $(RT_{60},D_{50})$. Actually, the characteristic function is computed for $\frac{\pi}{2}$ frequencies and stored in a vector of $\mathbb{R}^p$ where the first $\frac{p}{4}$ components are the real parts of $\phi_\rho(f)$ and the last $\frac{p}{4}$ components are their imaginary parts. The range of the frequencies and their sampling are chosen such that $|\phi_\rho(f)|$ lies between almost 0.1 and 1.

When a reverberant signal is recorded, we compute its energy decay rate distribution and the corresponding characteristic function, which is the feature vector. The output vector $\hat{y}$, which carries the room acoustic parameters, is estimated with the Nadaraya-Watson estimator (15).

### 4. PERFORMANCE EVALUATION

To evaluate our method we use anechoic speech signals of 25 English speakers (12 males, 12 females, one child; around 5 min each) from the TSP Speech database [24]. Our RIR database is composed of 1015 responses; among them 455 synthetic responses generated from the Fast Image-Source Method [25], with the RT$_{60}$ ranging from 0.1 s to 2.0 s and the source-microphone distance ranging from 0.1 m to 10 m. We add 560 real RIRs, from different open-access databases (Aachen Impulse Response [26], MARDY [27], OpenAIR [28], Queen-Mary [29]) with reverberation times ranging from 0.3 s to 8 s and C$_{50}$ from -10 dB to 25 dB. They can be observed in the (RT$_{60},C_{50}$) plane in Figure 1.

Reverberant speech signals are obtained by convolving each RIR with each anechoic speech and the energy decay rates are computed using frames of 32 ms with 50% overlap. The obtained distributions are centered and reduced, then the characteristic functions are computed for angular frequencies from 0 to 0.4 with 0.001 increments. Finally, we compute the average over the 25 different speakers, for each RIR.

We test our method in a 7-fold approach; we randomly split our corpus of characteristic functions in 7 subcorpora (fold) and successively estimate the room acoustic parameters of a fold, using the remaining 6 as dataset. The prediction error is then the average of errors obtained over the 7 tests. We use a Gaussian kernel with a bandwidth of $\lambda = 5 \cdot 10^{-4}$:

$$K_\lambda(x,x_i) = \frac{1}{\lambda} e^{-\frac{|x-x_i|^2}{\lambda^2}}. \quad \text{(16)}$$
5. RESULTS

We compute two kinds of errors, the relative error $e_{rel}$ (%) and the root mean square error $e_{rms}$ (same unit as the data):

$$e_{rel} = 100 \times \frac{1}{N} \sum_{i=1}^{N} \frac{x_i - \hat{x}_i}{x_i}$$

$$e_{rms} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}{N}}$$

with $\hat{x}_i$ the estimation of $x_i$ in a fold of $N$ tests. However, we will not use the relative errors when dealing with the $C_{50}$ since some are equals to zero.

5.1. Synthetic and real RIRs

We tested our method on a corpus made of both synthetic and real RIRs. The errors are then averaged over the 7 folds and presented in Table 1.

Figures 2 and 3 show the deviation between true and estimated parameters, for two different folds. The relative errors are around 10 % and 15 %, on a corpus including high reverberation times (up to 8 s).

However, the other methods found in the literature are tested for $RT_{60}$ up to 1 s or 2 s. Then, if we reduce our corpus to RIRs whose reverberation time does not exceed 2 s (usual reverberant environments) we obtain better results, summarized in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>$RT_{60} \leq 2s$</th>
<th>$RT_{60} \leq 1s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_{rms}$</td>
<td>180ms 9.3%</td>
<td>75ms 7.8%</td>
</tr>
<tr>
<td>$e_{rel}$</td>
<td>0.11dB 9.8%</td>
<td>0.07dB 11.8%</td>
</tr>
</tbody>
</table>

Table 2: Estimation error on the low $RT_{60}$ corpus, synthetic and real RIRs

With these lower reverberant environments estimation errors are lower, less than 10 % relative error for $RT_{60}$ and $D_{50}$, less than 1 dB of root mean square error for $C_{50}$ (which varies between 0 dB and 25 dB). Then we can compare our results to the performance of state of the art methods [19].

5.2. Comparison

We compare our results for $RT_{60}$ estimation to the ones provided in [19], which tests three methods of the literature (Modulation Energy Ratio (MER), Maximum Likelihood (ML), Spectral Decay Distributions (SDD)) on a database made of simulated and real RIRs. Even if we use the same simulation method for synthetic RIRs, and include the same real RIR database (AIR) in our corpus, we do not use the same corpus and then this comparison is intended to give an idea of the relative
Table 1: Estimation errors

<table>
<thead>
<tr>
<th></th>
<th>All RIRs</th>
<th>Synth.</th>
<th>Real</th>
</tr>
</thead>
<tbody>
<tr>
<td>$RT_{60}$</td>
<td>$e_{\text{rms}}$</td>
<td>$e_{\text{rel}}$</td>
<td>$e_{\text{rms}}$</td>
</tr>
<tr>
<td></td>
<td>496 ms</td>
<td>10.4%</td>
<td>143 ms</td>
</tr>
<tr>
<td>$D_{50}$</td>
<td>0.42 dB</td>
<td>14.3%</td>
<td>0.09 dB</td>
</tr>
<tr>
<td>$C_{50}$</td>
<td>1.07 dB</td>
<td></td>
<td>0.73 dB</td>
</tr>
</tbody>
</table>

Table 3: Comparison for $RT_{60}$ prediction

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20 %</td>
<td>15 %</td>
<td>10 %</td>
<td>7.8 %</td>
</tr>
</tbody>
</table>

Table 4: Comparison for $C_{50}$ prediction

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9.05 dB</td>
<td>5.52 dB</td>
<td>4.81 dB</td>
</tr>
</tbody>
</table>

We found in [15] a way of comparing the $C_{50}$ values. In this paper, the authors test their method (called NIRA) on the real RIRs database MARDY [27] and give the corresponding root mean square error. They also confront their results to the one obtained with their implementation of [8], used as a baseline. We then set a training corpus with all the characteristic functions obtained from our simulated and real RIRs, excluding the ones corresponding to the MARDY database. Then, we estimate the $C_{50}$ of the MARDY database with our method and this training dataset. For noise-free reverberant speech, the baseline obtains 9.05 dB RMS error, NIRA obtains 5.52 dB and our method 4.81 dB as we can see in Table 4.

6. CONCLUSION
As reverberation is determined by static properties of the room (resulting in the reverberation time) and the source-microphone distance (resulting in the clarity or definition index), we have presented a method to blindly jointly estimate these parameters. The distribution of the energy decay rates of a reverberant speech, described by its characteristic function, serves to perform kernel regression on a training dataset, with a Gaussian kernel and a low bandwidth. We tested our method in a 7-fold approach, with simulated and real room impulse responses and obtained a mean of 10.4 % relative error for $RT_{60}$ estimation, 14.3 % for $D_{50}$ estimation, 1.07 dB root mean square error for $C_{50}$ estimation. Compared to the available results from the literature, our method slightly outperforms the reverberation time estimation. There is no other method to blindly estimate the $D_{50}$ so we only compared the $C_{50}$ estimator to two state of the art methods, that were outperformed.

Future work will focus on the estimation of these parameters in noisy environments and introduce sparsity constraints in the feature vector, selecting the best frequencies of the characteristic function to be used.

7. REFERENCES


