SPARSE MODELING OF THE EARLY PART OF NOISY ROOM IMPULSE RESPONSES WITH SPARSE BAYESIAN LEARNING

Maozhong Fu[1,2]*, Jesper Rindom Jensen*[2]*, Yuhan Li*[1,2]*, and Mads Græsbøll Christensen*[2]

[1]School of Information Science and Engineering, Xiamen University, 361005 Xiamen, China
{maozhongfu,yuhanlily}@gmail.com
[2]Audio Analysis Lab, CREATE, Aalborg University, 9000 Aalborg, Denmark
{jrj, mgc}@create.aau.dk

ABSTRACT

A model of a room impulse response (RIR) is useful for a wide range of applications. Typically, the early part of a RIR is sparse, and its sparse structure allows for accurate and simple modeling of the RIR. The existing $\ell_p(0 < p \leq 1)$-norm-based methods suffer from the sensitivity to the user-selected regularization parameters or a high computational burden. In this work, we propose to reconstruct the sparse model for the early part of RIRs with sparse Bayesian learning (SBL). Under the framework of SBL, the proposed method can adaptively learn the optimal hyper-parameters from data at a low computational cost. Experiment results show that the proposed method has advantages in terms of noise robustness, reconstruction sparsity, and computational efficiency compared to the existing methods.

***Index Terms*—** Room impulse response, sparse modeling, sparse Bayesian learning.

1. INTRODUCTION

A measured room impulse response (RIR) characterizes the linear time-invariant system defined by the room, the source, and the receiver. A RIR are usually considered to comprise three parts: direct sound, early reflections, and late reverberation. Among them, the early reflections contain the first-order echoes which provide the information of the room geometry [1] and can contribute to the speech intelligibility [2]. Therefore, obtaining the model of the early part of RIRs is a fundamental step for various applications such as room geometry reconstruction [3, 4, 5], dereverberation [6, 7], augmented and virtual reality applications [8], and robot audition [9].

RIRs are typically modeled as long finite impulse response (FIR) filters which may have thousands of coefficients [10]. In order to reduce the dimensionality of the model, the ℓ_1-norm-based methods are used since they are sparsity-promoting [11, 12]. However, the performance of these approaches depends significantly on the user's choice of the regularization parameters. There are two common ways, including grid-search [13] and cross-validation [14] to choose the appropriate regularization parameters in real-world applications. But grid search is computation-heavy, and cross-validation requires suitable training data which is not always available. Modeling of RIRs can also be implemented as infinite impulse response (IIR) filters. By using orthonormal basis functions (OBFs) models to describe RIRs as particular fixed-poles IIR filters, a sparse approximation of an RIR can be found. [15]. Unfortunately, this method tries to fit both RIR and noise, therefore it may not work well in low signal-to-noise ratio (SNR) cases.

This work was supported by China Scholarship Council.

In this paper, we aim at improving the efficiency and sparsity of the modeling of the early reflections under the sparse Bayesian learning framework in noisy condition. Here, we improve that approach by greatly reduce the coefficient number of . The sparse nature of the early part of RIRs is effectively utilized by SBL to promote the sparsity of the model and suppress the noise in the model. The data-driven hyper-parameter learning adopted by SBL also potentially improves robustness for the proposed method in real-world applications. Moreover, due to the fast convergence of SBL, the proposed method is computationally efficient. This work is a follow-up of [16], where SBL is used to reconstruct RIRs for time-of-arriving (TOA) estimation. Here, we improve that approach by reducing the coefficient number of the final model and adopting new update rules with faster convergence.

Throughout this paper, we use the following notation. Matrices and vectors are represented by uppercase and lowercase letters, respectively. Vectors are column vectors. For a vector $\mathbf{a}$, $[\mathbf{a}]_n$ denotes the n-th entry of $\mathbf{a}$, and $\mathrm{diag}(\mathbf{a})$ denotes a diagonal matrix whose diagonal entries are the elements of $\mathbf{a}$. For a matrix $\mathbf{A}$, $\mathrm{tr}(\mathbf{A})$ denotes the trace of $\mathbf{A}$. For a sequence $\mathcal{S}$, $\{\mathcal{S}\}_n$ denotes the n-th entry of $\mathcal{S}$. For a sequence $\mathcal{S}$ and a vector $\mathbf{a}$, $[\mathbf{a}]_{\mathcal{S}}$ is a $K \times 1$ vector consisting of the K entries of $\mathbf{a}$ indexed by $\mathcal{S}$. For a sequence $\mathcal{S}$ and a matrix $\mathbf{A}$, $[\mathbf{A}]_{\mathcal{S}}$ is a $N \times K$ matrix consisting of the K columns of $\mathbf{A}$ indexed by $\mathcal{S}$. The $N \times N$ identity matrix is denoted by $\mathbf{I}^{(N)}$. The superscripts $(\cdot)^*$, $(\cdot)^T$, and $(\cdot)^+$ denote the conjugate, transpose, and Moore-Penrose pseudo-inverse operators, respectively. The operators $*$ and $\odot$ denote the convolution and the Hadamard product, respectively. The notation $\lfloor\cdot\rfloor$ denotes the floor function. The notation $\mathbb{E}(x)$ denotes the expectation expectation of a random variable x. The absolute value of a variable y is denoted by $|y|$. The ℓ_p-norm of a vector $\mathbf{a} \in \mathbb{R}^N$ is defined as $\|\mathbf{a}\|_p = \left(\sum_{n=1}^{N} |[\mathbf{a}]_n|^p\right)^{1/p}$.

2. SIGNAL MODEL AND PROBLEM FORMULATION

Under ideal conditions, the received signal in a reverberate environment can be mathematically expressed as $r(t) = h(t) * s(t) + w(t)$, where $h(t)$ is the RIR, $s(t)$ is the source signal, and $w(t)$ is the additive noise. Assume that $h(t)$ consists of I scaled and time-shifted Dirac impulse functions. The RIR can then be expressed as

$$h(t) = \sum_{i=1}^{I} \alpha_i \delta(t - \tau_i), \tag{1}$$

where α_i and τ_i are the attenuation coefficient and the time-delay of the i-th ($i \in [1, I]$) component, respectively. The sampled RIR

measurements in the discrete-time are given by

$$r((n-1)T_s) = \sum_{i=1}^{I} \alpha_i s((n-1)T_s - \tau_i) + w((n-1)T_s), \quad (2)$$

where $n \in [1, N]$ is the time index, $T_s = 1/F_s$ is the sampling interval, and F_s is the sampling frequency. The source signal $s(t)$ is usually modeled as a modulated Gaussian pulse (MGP) in the applications of RIR estimation [11, 17],

$$s(t) = \exp(-\theta^2 t^2)\cos(2\pi\xi t), \quad (3)$$

where θ and ξ are two parameters controlling the pulse shape. The time-delays in $h(t)$ may not be well approximated by multiples of T_s. To deal with fractional delays, we quantize the time-delay domain as $\tau = (m-1)T_\Delta$, where $T_\Delta = T_s/Q$ $(Q > 1 | Q \in \mathbb{N})$ is the quantization step size, $m \in [1, M]$ is the quantization index, and $M = QN$. Assume Q is large enough. In this case, $h(t)$ can be well approximated by M components as

$$h(t) = \sum_{m=1}^{M} \beta_m \delta(t - (m-1)T_\Delta), \quad (4)$$

where β_m is the attenuation coefficient of the m-th component with time-delay $(m-1)T_\Delta$. The number of the non-zero β_m is I, and the non-zero β_m correspond to the the attenuation coefficient α_i in (1). With (4), the RIR measurements in (2) become

$$r((n-1)T_s) = \sum_{m=1}^{M} \beta_i s((n-1)T_s - (m-1)T_\Delta) + w((n-1)T_s). \quad (5)$$

Equation (5) can be written in vector notation as

$$\mathbf{y} = \mathbf{\Phi}\mathbf{x} + \mathbf{w}, \quad (6)$$

where $\mathbf{y} = [r(0), r(T_s), \ldots, r((N-1)T_s)]^T$ is the vector of the RIR measurements, $\mathbf{x} = [\beta_1, \beta_2, \ldots, \beta_M]^T$ is the vector of the RIR coefficients,

$$\mathbf{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_M]^T \quad (7)$$

is the $N \times M$ matrix with the m-th column being the vector $\boldsymbol{\phi}_m = [s(-(m-1)T_\Delta), s(T_s - (m-1)T_\Delta), \ldots, s((N-1)T_s - (m-1)T_\Delta)]$, and $\mathbf{w} = [w(0), w(T_s), \ldots, w((N-1)T_s)]^T$ is the vector of the additive noise. Since the early part of RIRs typically is sparse, i.e., $\mathbf{x}$ is sparse, the estimation of the RIR model can be converted into the solving a sparse modeling problem [18].

3. PROPOSED METHOD

The system of linear equations in (6) typically is underdetermined, and the regularization-based methods which require user parameters are commonly used to find a solution. SBL avoids the choice of user parameters and estimates $\mathbf{x}$ by solving a probabilistic parameter estimation problem.

SBL regards $\mathbf{x}$ as a vector of stochastic variables. Thus, the estimate of $\mathbf{x}$ can be found by maximizing the posterior which can be given via Bayes's theorem as

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}, \quad (8)$$

where $p(\mathbf{x}|\mathbf{y})$ is the posterior, $p(\mathbf{y}|\mathbf{x})$ is the likelihood, $p(\mathbf{y})$ is the evidence, and $p(\mathbf{x})$ is the prior. To enforce the sparsity of $\mathbf{x}$, SBL adopts the zero-mean Gaussian distribution as the prior [19, 20], i.e.,

$$p(\mathbf{x}; \boldsymbol{\gamma}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{\Gamma}), \quad (9)$$

where $\mathbf{\Gamma} = \text{diag}(\boldsymbol{\gamma})$ and $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_M]^T$ is the vector of hyper-parameters. The hyper-parameter γ_m controls the strength of the prior over the m-th component of $\mathbf{x}$. When $\gamma_m = 0$, the m-th component of $\mathbf{x}$ equal to 0 with probability 1, which imposes sparsity constraints on $\mathbf{x}$.

We assume that $\mathbf{w}$ follows a zero-mean Gaussian distribution such that $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \lambda\mathbf{I}^{(N)})$. Therefore, the likelihood for $\mathbf{y}$ given $\mathbf{x}$ is

$$p(\mathbf{y}|\mathbf{x}; \lambda) = \mathcal{N}(\mathbf{y}|\mathbf{\Phi}\mathbf{x}, \lambda\mathbf{I}^{(N)}). \quad (10)$$

The evidence is the product of the prior (9) and likelihood (10) integrated over $\mathbf{x}$, which can be expressed as

$$\begin{aligned} p(\mathbf{y}; \boldsymbol{\gamma}, \lambda) &= \int p(\mathbf{y}|\mathbf{x}; \lambda)p(\mathbf{x}; \boldsymbol{\gamma})d\mathbf{x} = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{C}) \\ &= (2\pi)^{-N/2}|\mathbf{C}|^{-1/2}\exp\left(-\frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}\right), \end{aligned} \quad (11)$$

where

$$\mathbf{C} = \mathbb{E}(\mathbf{y}\mathbf{y}^T) = \lambda\mathbf{I}^{(N)} + \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T. \quad (12)$$

Substituting the prior (9) and likelihood (10) into Bayes' theorem, the posterior of $\mathbf{x}$ conditioned on $\boldsymbol{\gamma}$ and λ can be expressed as

$$p(\mathbf{x}|\mathbf{y}; \boldsymbol{\gamma}, \lambda) = \frac{p(\mathbf{y}|\mathbf{x}; \lambda)p(\mathbf{x}; \boldsymbol{\gamma})}{p(\mathbf{y}; \boldsymbol{\gamma}, \lambda)} \propto \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}), \quad (13)$$

where the posterior mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma}$ are

$$\boldsymbol{\mu} = \mathbf{\Gamma}\mathbf{\Phi}^T\mathbf{C}^{-1}\mathbf{y} \quad (14)$$

and

$$\mathbf{\Sigma} = \mathbf{\Gamma} - \mathbf{\Gamma}\mathbf{\Phi}^T\mathbf{C}\mathbf{\Phi}\mathbf{\Gamma}, \quad (15)$$

respectively. Since the posterior follows a Gaussian distribution, the maximum-a-posteriori (MAP) estimate of $\mathbf{x}$ is the posterior mean. Therefore, given the estimates of $\mathbf{\Gamma}$ and λ, the MAP estimate of $\mathbf{x}$ is

$$\hat{\mathbf{x}} = \boldsymbol{\mu} = \mathbf{\Gamma}\mathbf{\Phi}^T\mathbf{C}^{-1}\mathbf{y}. \quad (16)$$

The reconstructed RIR measurements can then be obtained as

$$\hat{\mathbf{y}} = \mathbf{\Phi}\hat{\mathbf{x}}. \quad (17)$$

The unknown $\boldsymbol{\gamma}$ and λ can be estimated from the evidence as described next. Taking the logarithm of (11) and keeping only the terms that are dependent on $\boldsymbol{\gamma}$ and λ, we can compute the log marginal likelihood from the evidence as

$$\mathcal{L}(\boldsymbol{\gamma}, \lambda) = -\frac{1}{2}\left(\log|\mathbf{C}| - \mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}\right). \quad (18)$$

After taking the derivative of (18) with respect to $\boldsymbol{\gamma}$ and setting the result to zero, the learning rule for $\boldsymbol{\gamma}$ is obtained as [21]

$$\boldsymbol{\gamma} \leftarrow \frac{\|\boldsymbol{\mu}\|_2}{\sqrt{((\mathbf{\Phi}^T\mathbf{C}) \odot \mathbf{\Phi}^T)\mathbf{J}^{(N)}}}, \quad (19)$$

where $\mathbf{J}^{(N)} = [1, 1, ..., 1]^T$ is a vector of N ones. With the updated $\boldsymbol{\gamma}$, a stochastic maximum likelihood-based learning rule of λ can be defined as [21]

$$\lambda \leftarrow \frac{1}{N-K}\,\text{tr}\left(\left(\mathbf{I}^{(N)} - [\mathbf{\Phi}]_{\mathcal{Q}}[\mathbf{\Phi}^+]_{\mathcal{Q}}\right)\mathbf{S}_y\right), \quad (20)$$

where $\mathbf{S}_y = \mathbf{y}\mathbf{y}^T$ is the covariance matrix of the measurements,

$$\mathcal{Q} = \{q \in \mathbb{N} | [\boldsymbol{\gamma}]_q \geq \boldsymbol{\gamma}^{(K)}\} \quad (21)$$

Algorithm 1 Proposed Method

Input: $\mathbf{\Phi}$, $\mathbf{y}$, K, ϵ_{stop}, N_{iter}, T_d
Output: $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$
procedure RIR-SBL($\mathbf{\Phi}, \mathbf{y}, K, \epsilon_{\text{stop}}, N_{\text{iter}}, T_d$)
  $\lambda \leftarrow 1, \gamma \leftarrow \mathbf{I}^{(M)}, \epsilon \leftarrow \infty, n_{\text{iter}} \leftarrow 1$
  $\tau_d \leftarrow T_s \times \arg\max_n |[\mathbf{y}]_n|$ ▷ (23)
  $\mathbf{d} \leftarrow [\mathbf{\Omega} \odot \mathbf{y}]_{\mathfrak{N}}$ ▷ (24)
  $\{\hat{\theta}, \hat{\xi}, \hat{\tau_d}\} \leftarrow \arg\min_{\theta,\xi,\tau_d} \frac{1}{N_d}(\mathbf{d}-\hat{\mathbf{d}})^T(\mathbf{d}-\hat{\mathbf{d}})$ ▷ (26)
  $\mathbf{\Phi} \leftarrow [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_M]^T$ ▷ (7)
  while $\epsilon < \epsilon_{\text{stop}}$ and $n_{\text{iter}} \leq N_{\text{iter}}$ **do**
    $\mathbf{C} \leftarrow \lambda \mathbf{I}^{(N)} + \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T$ ▷ (12)
    $\boldsymbol{\mu} \leftarrow \mathbf{\Gamma}\mathbf{\Phi}^T\mathbf{C}^{-1}\mathbf{y}$ ▷ (14)
    $\boldsymbol{\gamma} \leftarrow \frac{\|\boldsymbol{\mu}\|_2}{\sqrt{((\mathbf{\Phi}^T\mathbf{C})\odot\mathbf{\Phi}^T)\mathbf{J}^{(N)}}}$ ▷ (19)
    $\mathcal{Q} \leftarrow \{q \in \mathbb{N} | \gamma_q \geq \gamma^{(K)}\}$ ▷ (21)
    $\lambda \leftarrow \frac{1}{N-K}\operatorname{tr}\left(\left(\mathbf{I}^{(N)} - [\mathbf{\Phi}]_{\mathcal{Q}}[\mathbf{\Phi}^+]_{\mathcal{Q}}\right)\mathbf{S}_y\right)$ ▷ (20)
    $\epsilon \leftarrow \|\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\|_1 / \|\boldsymbol{\gamma}\|_1$ ▷ (22)
    $n_{\text{iter}} \leftarrow n_{\text{iter}} + 1$
  end while
  $\hat{\mathbf{x}} \leftarrow \boldsymbol{\mu}$ ▷ (16)
  $\hat{\mathbf{y}} \leftarrow \mathbf{\Phi}\hat{\mathbf{x}}$ ▷ (17)
end procedure

is the set of indices which indicates the positions of the K largest elements of $\boldsymbol{\gamma}$, and $\boldsymbol{\gamma}^{(K)}$ is the K-th largest elements of $\boldsymbol{\gamma}$.

The updates for $\boldsymbol{\gamma}$ and λ should be carried out iteratively until convergence. The convergence criterion of the iteration is defined as

$$\epsilon = \|\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\|_1 / \|\boldsymbol{\gamma}\|_1 , \tag{22}$$

where $\hat{\boldsymbol{\gamma}}$ are the estimated hyper-parameters in the last iteration. The iteration terminates when ϵ is smaller than a specified value ϵ_{stop} or the maximum number of iterations N_{iter} is reached. It can be noted that most elements of $\boldsymbol{\gamma}$ may be close to 0 when convergence is reached, which indirectly drives most elements of $\mathbf{x}$ to be 0.

4. DICTIONARY BUILDING

In practice, the source signal $s(t)$ may be unknown. In such cases, we need to estimate the parameters θ and ξ of the adopted MGP model to build the dictionary.

The direct sound provides useful information for the estimation of θ and ξ. Assume the direct sound has the highest amplitude in the measurements. A coarse estimate of the time-delay of the direct sound can be firstly obtained by finding the maximum of the measurements,

$$\tau_d = T_s \times \arg\max_n |[\mathbf{y}]_n| . \tag{23}$$

Then the component of the direct sound is extracted by adding a rectangular window to $\mathbf{y}$,

$$\mathbf{d} = [\boldsymbol{\omega} \odot \mathbf{y}]_{\mathfrak{N}}, \tag{24}$$

where $\mathbf{d}$ is the vector of the measurements of the direct sound, $\boldsymbol{\omega} = [\omega_1, \omega_2, \ldots, \omega_N]^T$ is the vector of the rectangular window, $\omega_n = \text{rect}\left(((n-1)T_s - \tau_d)/T_d\right)$, T_d is the window length of the direct sound,

$$\text{rect}(u) = \begin{cases} 1, & \text{for } |u| < 0.5 \\ 0, & \text{otherwise,} \end{cases} \tag{25}$$

$\mathfrak{N} = \{n | \omega_n = 1\}$ is an increasing sequence that consists of the indices of the measurements within the rectangular window, and N_d is the length of $\mathfrak{N}$. After that, the estimates of θ and ξ can be found by minimizing the mean squared error (MSE) between the reconstructed direct sound and the measured direct sound in the window, i.e.,

$$\{\hat{\theta}, \hat{\xi}, \hat{\tau_d}\} = \arg\min_{\theta,\xi,\tau_d} \frac{1}{N_d}(\mathbf{d}-\hat{\mathbf{d}})^T(\mathbf{d}-\hat{\mathbf{d}}), \tag{26}$$

where $\hat{\mathbf{d}} = [s((\{\mathfrak{N}\}_1 - 1)T_s - \tau_d), s((\{\mathfrak{N}\}_2 - 1)T_s - \tau_d), \ldots, s((\{\mathfrak{N}\}_{N_d} - 1)T_s - \tau_d)]^T$ is the vector of the reconstructed direct sound. The dictionary is finally built by substituting $\hat{\theta}$ and $\hat{\xi}$ into (7). Algorithm 1 summarizes the procedures of the proposed method.

Table 1: Parameters settings of the simulations.

Symbol	Parameter	Value
F_s	Sampling Frequency	48 000 Hz
V	Room Geometry	$(3.32, 4.83, 2.95)$ m
T60	Reverberation Time	0.4 s
Tx	Source Position	$(1, 2.9, 0.5)$ m
Rx	Microphone Position	$(0.5, 1, 0.5)$ m
D	Microphone Directivity	$(0, 1, 0)$
P	Microphone Pattern	Cardioid
SNR_{in}	Signal-to-Noise Ratio	5 dB

5. RESULTS

The experiments are carried out with simulated and real-world RIR measurements. The simulated noiseless RIR measurements are generated by the image method using the gpuRIR library [22], and zero mean white Gaussian noise is added to form the final noisy RIR measurements. The parameter settings for the simulations are set as Table 1, if not specified otherwise. For the simulated RIR measurements, the input SNR is defined as $\text{SNR}_{\text{in}} = 10\log_{10}\|\mathbf{\Phi}\mathbf{x}\|_2^2 / \|\mathbf{w}\|_2^2$ (dB) [23]. Similarly, the output SNR of the obtained RIR measurements is defined as $\text{SNR}_{\text{out}} = 10\log_{10}\|\mathbf{\Phi}\mathbf{x}\|_2^2 / \|\hat{\mathbf{y}} - \mathbf{\Phi}\mathbf{x}\|_2^2$ (dB). The real-world RIR measurements are taken from the acoustic characterization of environments (ACE) database [24]. In the implementation of the proposed method, $Q = 10$, $K = 1$, $N_{\text{iter}} = 1000$, $\epsilon_{\text{stop}} = 10^{-6}$, and $T_d = 0.4$ ms are adopted. The simulated and real-world RIR measurements all are truncated to $N = 400$ samples which start from the 10 samples before the first strong peak (the direct sound) and normalized to 1 by the absolute value of the direct sound amplitude.

5.1. Evaluation results with the simulated RIR measurements

The results for the sparse modeling of the early part of RIRs at a high SNR ($\text{SNR}_{\text{in}} = 10$ dB) and a low SNR ($\text{SNR}_{\text{in}} = 0$ dB) are shown in Fig. 1(a) and (b), respectively. As it can be observed, the proposed method, referred to as RIR-SBL, exhibits higher performance in noise reduction both for the high SNR and low SNR cases than the LASSO-based method in [11], referred to as RIR-LASSO. Especially in the low SNR case, RIR-SBL shows less noise sensitiveness to the noise than RIR-LASSO in the time interval [3,7] ms.

Fig. 2 plots the curves of the averaged output SNRs with different input SNRs. The averaged output SNRs are calculated over 10 Monte Carlo trials. It can be seen that RIR-SBL provides an improvement in the output SNR compared to RIR-LASSO.

Fig. 3 presents the histogram of the reconstructed RIR measurements. It can be observed that the reconstructed RIR measurements of RIR-SBL are more concentrated around zero than that of RIR-

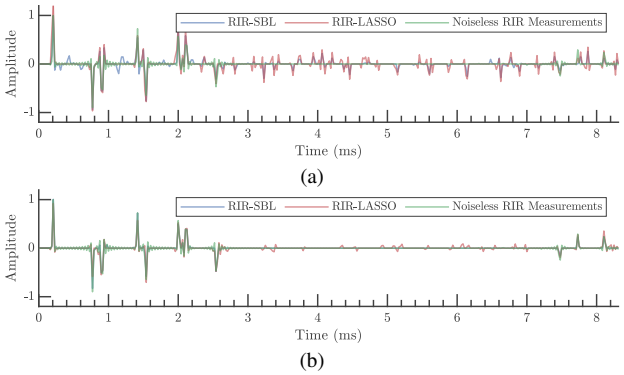

Fig. 1: Results for the sparse modeling of the early part of RIRs at different input SNRs. (a) $\text{SNR}_{\text{in}} = 0$ dB. (b) $\text{SNR}_{\text{in}} = 10$ dB.

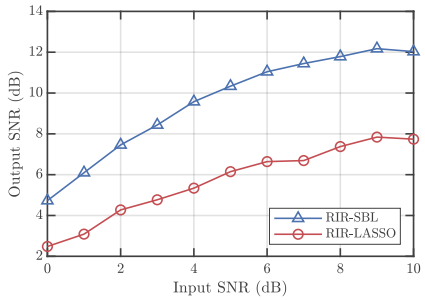

Fig. 2: Averaged output SNRs at different input SNRs.

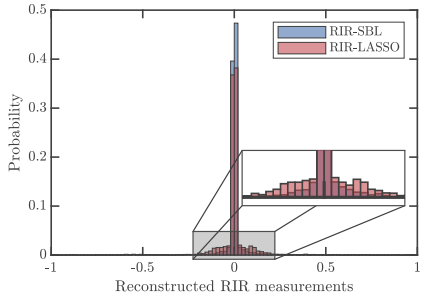

Fig. 3: Histogram of the reconstructed RIR measurements.

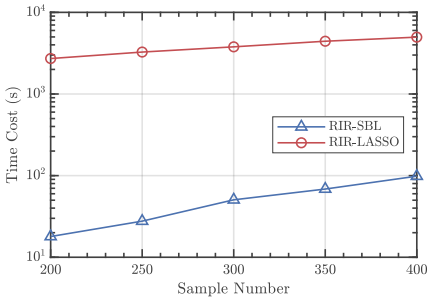

Fig. 4: Computational cost versus various sample numbers.

LASSO, which implies that the reconstruct RIR measurements of RIR-SBL is sparser than that of RIR-LASSO.

Fig. 4 shows the comparison of computational cost between RIR-SBL and RIR-LASSO. This experiment is executed on a Intel(R) Xeon(R) E5-2678 CPU 2.50 GHz with Windows Server 2016 and MATLAB 2019a. Each method is run three times to obtain the execution time τ_1, τ_2, and τ_3. The corresponding computational cost is then $\tau = \min(\tau_1, \tau_2, \tau_3)$ which is assumed to be free from the influence of other system processes [25]. It can be observed from Fig. 4 that RIR-SBL requires much less computation than RIR-LASSO.

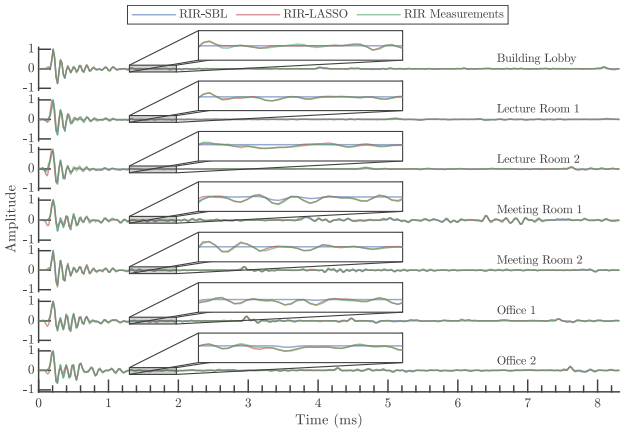

Fig. 5: Results for the sparse modeling of the early part of real-world RIR measurements.

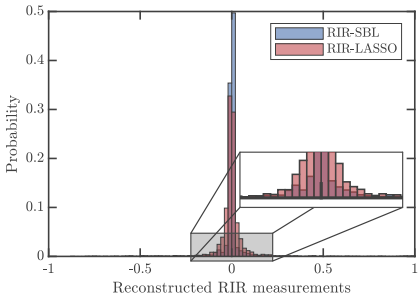

Fig. 6: Histogram of the reconstructed real-world RIR measurements.

5.2. Evaluation results with the real-world RIR measurements

Fig. 5 shows the results for the sparse modeling of real-world RIR measurements for different rooms. It can be seen that RIR-SBL has more reconstructed RIR measurements close to zero than RIR-LASSO in all seven different rooms, which implies that RIR-SBL performs better in noise reduction in real-world RIR processing.

Fig. 6 shows the histogram of the reconstructed real-world RIR measurements. The RIR measurements shown in Fig. 5 are used to form the histogram. It is clear that the distribution of RIR-SBL has a narrower shape than that of RIR-LASSO. This result shows that RIR-SBL is more sparse-promoting than RIR-LASSO in real-world RIR processing.

6. CONCLUSIONS

In this work, a novel method for sparse modeling of the early part of RIRs is proposed. The proposed method is based on the sparse nature of early reflections and the ability of SBL to solve sparse modeling problems. Experiments show that the advantages of the proposed method include the following: a) The proposed method works well in noisy conditions and can improve the output SNR. b) The proposed method is computationally efficient. c) The sparse structure of the early reflections is effectively exploited to promote the sparsity of the model by the proposed method.

7. REFERENCES

[1] I. Dokmanic, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proceedings of the National Academy of Sciences*, vol. 110, no. 30, pp. 12186–12191, 2013.

[2] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *The Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3233–3244, 2003.

[3] Y. E. Baba, A. Walther, and E. A. P. Habets, "3D Room Geometry Inference Based on Room Impulse Response Stacks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 857–872, 2018.

[4] F. Antonacci, J. Filos, M. R. P. Thomas, E. A. P. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, "Inference of Room Geometry From Acoustic Impulse Responses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2683–2695, 2012.

[5] F. Ribeiro, D. Florencio, D. Ba, and C. Zhang, "Geometrically Constrained Room Modeling With Compact Microphone Arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1449–1460, 2012.

[6] D. Baby and H. Van hamme, "Joint Denoising and Dereverberation Using Exemplar-Based Sparse Representations and Decaying Norm Constraint," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 2024–2035, 2017.

[7] I. Kodrasi and S. Doclo, "Joint Dereverberation and Noise Reduction Based on Acoustic Multi-Channel Equalization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 680–693, 2016.

[8] O. Das, P. Calamia, and S. V. Amengual Gari, "Room Impulse Response Interpolation from a Sparse Set of Measurements Using a Modal Architecture," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 960–964.

[9] U. Saqib, A. Deleforge, and J. R. Jensen, "Detecting Acoustic Reflectors Using A Robot's Ego-Noise," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 466–470.

[10] H. Hacihabibouglu and Z. Cvetkovic, "Multichannel Dereverberation Theorems and Robustness Issues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 676–689, 2012.

[11] C. Papayiannis, C. Evers, and P. A. Naylor, "Sparse parametric modeling of the early part of acoustic impulse responses," in *2017 25th European Signal Processing Conference (EUSIPCO)*, Kos, Greece, 2017, pp. 678–682, IEEE.

[12] M. Crocco and A. Del Bue, "Room impulse response estimation by iterative weighted L1-norm," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1895–1899.

[13] E. Guizzo, T. Weyde, and J. B. Leveson, "Multi-Time-Scale Convolution for Emotion Recognition from Speech Audio Signals," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6489–6493.

[14] G. N. Lilis, D. Angelosante, and G. B. Giannakis, "Sound Field Reproduction using the Lasso," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 1902–1912, 2010.

[15] G. Vairetti, T. van Waterschoot, M. Moonen, M. Catrysse, and S. H. Jensen, "Sparse linear parametric modeling of room acoustics with Orthonormal Basis Functions," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 1–5.

[16] Z. Bai, J. R. Jensen, J. Sun, and M. G. Christensen, "A Sparse Bayesian Learning Based RIR Reconstruction Method for Acoustic Toa And DOA Estimation," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 393–397.

[17] P. Antsalo, A. Makivirta, V. Valimaki, T. Peltonen, and M. Karjalainen, "Estimation of Modal Decay Parameters from Noisy Response Measurements," in *Audio Engineering Society Convention 110*. 2001, Audio Engineering Society.

[18] I. Rish and G. Grabarnik, *Sparse Modeling: Theory, Algorithms, and Applications*, CRC Press, 1st edition edition, 2014.

[19] D. P. Wipf and B. D. Rao, "An Empirical Bayesian Strategy for Solving the Simultaneous Sparse Approximation Problem," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3704–3716, 2007.

[20] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, 2004.

[21] P. Gerstoft, C. F. Mecklenbrauker, A. Xenaki, and S. Nannuru, "Multisnapshot Sparse Bayesian Learning for DOA," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1469–1473, 2016.

[22] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.

[23] C. Huszty and S. Sakamoto, "A note on the definition of signal-to-noise ratio of room impulse responses," *Acoustical Science and Technology*, vol. 33, no. 2, pp. 117–120, 2012.

[24] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of Room Acoustic Parameters: The ACE Challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1681–1693, 2016.

[25] T. L. Jensen, J. K. Nielsen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "A Fast Algorithm for Maximum-Likelihood Estimation of Harmonic Chirp Parameters," *IEEE Transactions on Signal Processing*, vol. 65, no. 19, pp. 5137–5152, 2017.