Unified Variational and Physics-aware Model for Room Impulse Response Estimation

Louis Lalay, Mathieu Fontaine, Roland Badeau

Unified Variational and Physics-aware Model for Room Impulse Response Estimation

Louis Lalay[1], *Mathieu Fontaine*[1], *Roland Badeau*[1]

[1]Télécom Paris, Institut Polytechnique de Paris, LTCI, France

louis.lalay@telecom-paris.fr, mathieu.fontaine@telecom-paris.fr, roland.badeau@telecom-paris.fr

Abstract

Room impulse response estimation is essential for tasks like speech dereverberation, which improves automatic speech recognition. Most existing methods rely on either statistical signal processing or deep neural networks designed to replicate signal processing principles. However, combining statistical and physical modeling for RIR estimation remains largely unexplored. This paper[1] proposes a novel approach integrating both aspects through a theoretically grounded model. The RIR is decomposed into interpretable parameters: white Gaussian noise filtered by a frequency-dependent exponential decay (e.g. modeling wall absorption) and an autoregressive filter (e.g. modeling microphone response). A variational free-energy cost function enables practical parameter estimation. As a proof of concept, we show that given dry and reverberant speech signals, the proposed method outperforms classical deconvolution in noisy environments, as validated by objective metrics.

Index Terms: Reverberation, room impulse response, variational theory

1. Introduction

The estimation of the Room Impulse Response (RIR) is a crucial task with applications to room volume estimation [1], acoustic matching [2, 3], which enables transforming audio between different environments, and automatic speech recognition (ASR) [4, 5, 6]. Another key front-end task related to RIR estimation is dereverberation, which improves ASR performance. However, since deconvolution with a degraded RIR is highly sensitive, most dereverberation methods bypass explicit RIR estimation and instead directly recover the dry signal using signal processing [7], stochastic approaches [8, 9], or predominantly deep neural networks (DNNs) [10, 11, 12, 13].

RIR estimation remains nevertheless an active research field, serving as a cornerstone for the aforementioned applications. RIR estimation techniques have evolved over several decades, introducing various models of different types. One of the earliest models was proposed by Schröder in [14], where the RIR is represented as the product of a white Gaussian process and an exponential decay. Later, Polack [15] demonstrated that the Gaussian component varies slowly in the frequency domain, leading to more accurate RIR estimation in practice. While these models effectively capture late reverberation, they struggle to accurately represent early reflections. To address this limitation, Komatsu *et al.* [16] introduced a spatial-temporal Gaussian process model that accounts for both early and late reverberation. Other methods, such as those in [17, 18], estimate RIR by formulating optimization problems, incorporating optimal transport regularization [18] or using low-rank approximations [17]. Among deep neural network (DNN) approaches, generative-based models [3], signal processing-guided methods [19], and more compact architectures like encoder-decoder convolutional neural networks [20] have been explored. However, these models do not simultaneously capture spatial, temporal, and frequency information, despite their theoretical correlation. The work in [21] represents an initial attempt to introduce a spatiotemporal and frequency-dependent model. Its first application demonstrated that, given a real RIR, it enables accurate estimation of the reverberation time after 60dB of attenuation (RT_{60}) [22].

We propose in this paper a variational formulation using the physical model in [21] to perform RIR estimation [2]. In such a way, we derive a variational free-energy loss that will be used in order to estimate the model parameters of the variational distribution. In particular, the variational distribution is assumed to be a Gaussian whose mean is the targeted RIR. In this first proof of concept, we assume that both dry and reverberant signals are known, which can be seen as a subtask of acoustic matching. Our observed signal is however corrupted with additive noise to see the robustness of our approach. Our results show the clear robustness of the proposed technique compared to other classic deconvolution techniques [23] in terms of estimation of rever-

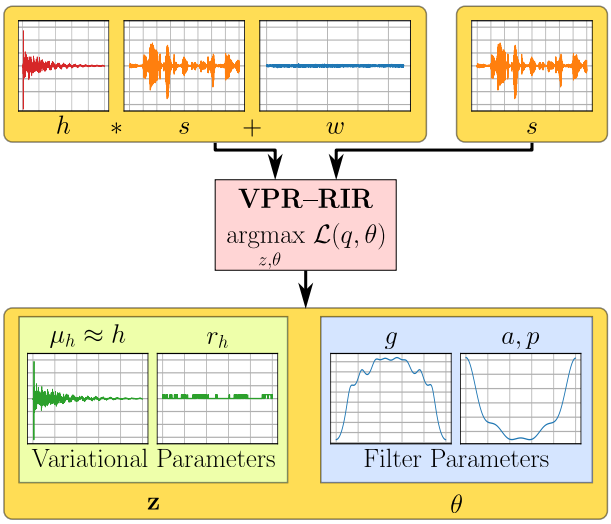

Figure 1: *Outline of the proposed variational and physics-aware reverberation model for room impulse response estimation (VPR-RIR)*

[1]This paper was submitted to Interspeech 2025

[2]The code is available at https://github.com/LouisLalay/VPR-RIR

beration characteristics (RT_{60}, RT_{30}, direct-to-reverberant ratio (DRR), energy decay relief (EDR), etc.).

2. Method

In this section we describe the model we chose for its accuracy in RIR description, then we present the mixing model considered for the experiments and finally the variational approach used to estimate the model parameters from the observations.

2.1. RIR and Mixing Model

Let $y \in \mathbb{R}^T$ be the measured signal, $s \in \mathbb{R}^{L_s}$ be the source signal and $h \in \mathbb{R}^{L_h}$ be the room impulse response of length T, L_s and L_h respectively, such that $T = L_h + L_s - 1$. We consider in the time domain the following mixing model:

$$y = s * h + w, \tag{1}$$

where $w \in \mathbb{R}^T$ is an additive white Gaussian noise of variance σ_w^2. We model the random vector h as in [22]:

$$h = G^{-1}E^{-1}P^{-1}\epsilon \tag{2}$$

where the components are defined as follows:

- $\epsilon[u] \sim \mathcal{N}(0, \sigma_\epsilon^2)$ for $u = 0, \dots, L_h - 1$ (i.i.d. white Gaussian noise);
- $G = \text{Toep}(g)$, $g \in \mathbb{R}^{L_g}$, lower triangular;
- $E = \text{Diag}(\{e^{au}\}_{u=0}^{L_h-1})$, $a \in [0, 1]$, diagonal;
- $P = \mathcal{P}_{L_h}(p)$ with $p_0 = 1$, $p \in \mathbb{R}^{L_p}$, lower triangular.

All matrices are square and of size $L_h \times L_h$ which is the length of the RIR in the number of samples. Toep(b), Diag(b) stand for the lower triangular Toeplitz and the diagonal matrix, respectively, generated using a vector b. $\mathcal{P}_{L_h}(p)$ is the space of the square matrices of size L_h generated by p. The matrix $\mathcal{P}_4(p)$ is for instance constructed as described below, with $L_h = 4, p = [p_0, p_1]$ where $p^{*n} = p * p^{*(n-1)}$ with $p^{*0} = \delta$ (where δ denotes the unit impulse at time 0) and $p^{*n}[:l]$ the truncated vector of the l^{th} first entries of p^{*n}.

$$\mathcal{P}_4([p_0, p_1]) = \begin{array}{c} \begin{matrix} p^{*0}[:4] & p^{*1}[:4] & p^{*2}[:4] & p^{*3}[:4] \end{matrix} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & p_0 & 0 & 0 \\ 0 & p_1 & p_0^2 & 0 \\ 0 & 0 & 2p_0p_1 & p_0^3 \end{pmatrix} \end{array}$$

The RIR model is based on physical assumptions, and it can be split into four parts. First, ϵ is the base Gaussian white noise that is filtered to obtain the RIR. Then the Toeplitz matrix G, parameterized by the vector g of length $L_g < L_h$, implements a filter that models the microphone response. The diagonal matrix E accounts for the average absorption of the room, parameterized by the absorption coefficient a. Finally, the matrix P accounts for the frequency-dependent absorption of the walls, parameterized by filter p of length $L_p < L_h$. Having 3 matrices to model the global filter, we can force $g_0 = 1$ and $p_0 = 1$ without losing generality and still benefitting from good properties in terms of inversion and determinant computations.

2.2. Parameters estimation

We adopt a variational approach to estimate the model parameters. Let $\theta = \{g, a, p, \sigma_\epsilon, \sigma_w\}$ represent the set of model parameters that we aim to estimate. The mean field target distribution for h that we aim to approximate is given by $q^*(h[u]) = \mathcal{N}(\mu_h[u], r_h[u])$. Let $z = \{\mu_h, r_h\}$ represent the optimal distribution parameters where μ_h, r_h are the vectors with the components $\mu_h[u], r_h[u]$ respectively. The variational free energy (VFE) associated with this problem is defined as [24]:

$$\mathcal{L}(q, \theta) = \mathbb{E}_{q(z)}\left[\ln \frac{\mathbb{P}(y, z \mid \theta, s)}{q(z)}\right] \tag{3}$$

By combining the stochastic model of the RIR in Eq. (2) and Eq. (1), we can get a closed form of Eq. (3):

$$\begin{aligned} \mathcal{L}(q, \theta) = &-\frac{1}{2}\left(T\ln(2\pi\sigma_w^2) + \frac{\mathbb{E}_{q(z)}\left[\|y - s * h\|_2^2\right]}{\sigma_w^2}\right) \\ &- \frac{1}{2}\left(L_h \ln(\sigma_\epsilon^2) - L_h(L_h - 1)a\right) \\ &- \frac{1}{2\sigma_\epsilon^2}\left(\|V\mu_h\|_2^2 + \text{Tr}\left(VR_hV^\top\right)\right) \\ &+ \frac{1}{2}\left(\sum_{u=0}^{L_h-1} \ln r_h[u] + L_h\right) \end{aligned} \tag{4}$$

where $.^\top$, Tr(.) are the transposition and trace operator respectively, $\|.\|_2$ stands for the Euclidean norm of a real vector, $V = PEG$, $R_h = \text{Diag}(r_h)$, and

$$\begin{aligned} \mathbb{E}_{q(z)}\left[||y - s * h||_2^2\right] =& ||y||_2^2 - 2\sum_{t=0}^{T-1} y[t](\mu_h * s)[t] \\ &+ \sum_{t=0}^{T-1}(r_h * s^2)[t] + (\mu_h * s)^2[t]. \end{aligned}$$

We want to maximize this function for z and θ. We can formulate this as an optimization problem:

$$(z, \theta) = \underset{z,\theta}{\text{argmax}}\ \mathcal{L}(q, \theta). \tag{5}$$

We can then proceed to the estimation of the parameters by gradient descent. In the experiments, we choose the Adam optimizer [25] to solve this problem. Note that a variational expectation-maximization algorithm [24] can also be adapted to the problem to retrieve the parameters, and is left for future work.

2.3. Consistency constraint on g

In theory, the filter g should cut low and high frequencies because it represents the response of the microphone, which is a physical band pass filter. However, our preliminary studies demonstrate that such a constraint was not respected along the iterations, leading sometimes to a poor RIR estimation. To solve this issue, we force 2 zeros at Nyquist and null frequencies in the transfer function of the filter to get a better estimation. The matrix G_{old} becomes $G = G_1G_0$ with $G_1 = G_{old}$ and $G_0 = \text{Toep}([1, 0, -1])$.

Moreover, to ensure consistency in the model, the filter g is stabilized to keep G^{-1} numerically stable. The stable version of g is obtained by placing all the unstable poles inside the unit circle. The change in global gain of the filter is not compensated.

The parameters are initialized with $PEG = I_{L_h}$, $R_h = I_{L_h}$, $\mu_h = \delta$, $\sigma_w = 1$, $\sigma_\epsilon = 1$. The normalization step ensures

g is stable, a is strictly positive and r_h is strictly positive for coherence with the model. We summarize the proposed iterative approach in Algorithm 1.

Algorithm 1 Proposed RIR estimation

Initialization: initialize θ, z
Number of iterations I
Adam optimizer for all parameters θ, z with $lr = 1e-3$
for $i = 1$ to I **do**
    Construct P, E, G using current θ
    Normalize the parameters
    Optimizer: zero grad
    Compute the loss: $-\mathcal{L}(q, \theta)$
    Backward propagation
    Update θ, z
end for
Return: θ, z

3. Experimental setup

This section outlines the data, metrics, and methods used in our evaluation. We conduct two tasks: the first one involves RIR estimation in different settings, and the second one focuses on the reconvolution of dry speech using the estimated RIR.

3.1. Datasets

Speech Dataset In all experiments, we use a single audio utterance, `87-8000`, randomly sampled from the LibriSpeech dataset [26], truncated at 2 s. The original sampling rate is 16 kHz, but for our evaluation, we down-sampled it to 8 kHz. Preliminary experiments conducted at both 16 kHz and 8 kHz yielded similar results, so we opted for the 8 kHz version for computational efficiency.

RIR Dataset To demonstrate the robustness of our algorithm for RIR estimation, we utilize both real (denoted "Real") and synthetic (denoted "Synthetic") RIRs with various RT_{60} values that we qualified hereafter as "Short" when $\text{RT}_{60} \in [0, 0.5]$ s and "Long" when $\text{RT}_{60} \in [1.0, 1.5]$ s. The synthetic reverb comes from HybridReverb2 [27] while the real ones are extracted from BUT ReverbDB dataset [28]. BUT ReverbDB contains RIRs from 8 different rooms with various microphone and loudspeaker positions. HybridReverb2 contains synthetic RIRs generated by an auralization software with 6 presets for the geometry of the room and multiple speaker positions. We consider 2000 RIR samples, corresponding to 250 ms. This choice is primarily due to the high computational cost of our system and, secondly, because we aimed to focus on estimating the early reflection part, which is the most challenging aspect of RIR estimation, rather than the late reverberation.

Noise Dataset As our model in Eq. (1) takes into account some Gaussian noise, we decided to add noise to the reverberant speech signal. We consider one easy task where the additive noise is a Gaussian with an SNR of 40 dB and a second one more challenging with 5 randomly chosen noise files coming from WHAMR! [29] dataset and an SNR of 20 dB. Concerning the WHAMR! additive noise, we only report results with the convolution of "Short" and "Real" RIR as the results remain the same in other settings.

3.2. Metrics and Task Description

RIR estimation We consider the mean absolute error between an estimated RIR $\tilde{h}$ and the ground-truth RIR h for various acoustic characteristics $m \in \{$C80, D50, DRR, EDC, EDR, RT30, RT60$\}$ that we denote Δ_m and which represents the clarity at 80 ms, definition at 50 ms, direct to reverberant ratio, energy decay curve, energy decay relief, reverberation time at 30 and 60 dB respectively. We moreover consider the mean-squared error (MSE).

Speech reconvolution For this task, we get an estimated RIR $\tilde{h}$ that we convolve with the dry signal s to get an approximate $\tilde{y} = \tilde{h} * s$. We then consider the Wide-band Perceptual Speech Quality Score (PESQ [30]) and the short-time objective intelligibility (STOI [31]) measure between y and $\tilde{y}$.

3.3. Methods

We consider in our experiments the proposed approach and two baseline methods:

- VPR-RIR: our proposed Variational and Physics-aware Reverberation model for Room Impulse Response. For the sake of clarity, we gather the parameters initialization and procedure in Algorithm 1.
- DEC: this technique is a *spectral deconvolution* where y, h, s are considered in the Fourier domain and $Y = \mathcal{F}(y)$, $S = \mathcal{F}(s)$ are the Fourier transforms of the signals padded to the same length. We then compute $h = \mathcal{F}^{-1}(Y/S)$, where $\mathcal{F}^{-1}$ is the inverse Fourier transform.
- CBF: We consider the Cross Band Filtering defined in [23] with the following parameters: $K = 1$ crossbands, $n_{fft} = 512$, 50% overlap and a Hann window.

4. Results & Discussions

4.1. Room Impulse Response Estimation

Tables 1 and 2 present the error results for the Gaussian noise and WHAMR! noise setups, respectively. From Table 1, we observe that the results remain relatively comparable regardless of whether the RIR is real, synthetic, short, or long. However, we note a significant standard deviation in Δ_{C80} for DEC and CBF in the synthetic case, while VPR-RIR exhibits a notable standard deviation in the short real case. It is important to highlight that this setting favours DEC and CBF, as the signal-to-noise ratio (SNR) in such conditions is high, making DEC an exact solution for this type of problem. Overall, these results confirm that VPR-RIR is competitive with state-of-the-art methods, regardless of the RIR type.

Next, we examine a more challenging scenario where real noise is added to the reverberant signal, with results shown in Table 2. In this case, VPR-RIR consistently outperforms DEC and CBF across all error metrics and standard deviations.

These initial experiments demonstrate that VPR-RIR has the potential to estimate an RIR from which acoustic characteristic parameters can be extracted. Future work may explore a wider range of SNR levels with realistic noise to further evaluate the capabilities of VPR-RIR. This improved performance is achieved at the expense of computational efficiency, though there remains ample room for optimization to reduce runtime.

4.2. Speech Reconvolution

Figure 2 presents boxplot results for PESQ and STOI scores of the speech reconvolution (see Section 3.2 for more details).

		Real		Synthetic	
		Short	Long	Short	Long
$\Delta_{C80}(\downarrow)$	VPR-RIR	0.360 ± 0.120	0.990 ± 0.570	0.680 ± 0.870	2.100 ± 0.230
	DEC	**0.140 ± 0.180**	0.026 ± 0.026	**0.270 ± 4.000**	**0.044 ± 0.078**
	CBF	0.500 ± 0.290	**0.150 ± 0.170**	0.790 ± 4.800	0.048 ± 0.110
$\Delta_{D50}(\downarrow)$	VPR-RIR	0.014 ± 0.037	0.049 ± 0.032	0.005 ± 0.005	0.096 ± 0.027
	DEC	0.003 ± 0.002	**0.002 ± 0.001**	**0.002 ± 0.001**	**0.002 ± 0.001**
	CBF	**0.010 ± 0.009**	0.008 ± 0.010	0.004 ± 0.002	0.003 ± 0.002
$\Delta_{DRR}(\downarrow)$	VPR-RIR	1.3e-4 ± 2.1e-2	5.3e-5 ± 2.0e-2	1.6e-2 ± 2.6e-1	1.8e-5 ± 1.9e-2
	DEC	**1.7e-5 ± 4.3e-3**	**1.4e-5 ± 2.0e-2**	**1.8e-5 ± 3.2e-1**	1.6e-5 ± 3.0e-1
	CBF	1.2e-4 ± 4.2e-3	3.9e-5 ± 2.0e-2	3.1e-5 ± 1.7e-1	**1.6e-5 ± 1.7e-1**
$\Delta_{EDC}(\downarrow)$	VPR-RIR	0.007 ± 0.024	0.029 ± 0.017	0.006 ± 0.003	0.052 ± 0.016
	DEC	**0.002 ± 0.001**	**0.001 ± 0.000**	**0.001 ± 0.000**	**0.001 ± 0.001**
	CBF	0.009 ± 0.004	0.006 ± 0.004	0.003 ± 0.000	0.002 ± 0.001
$\Delta_{EDR}(\downarrow)$	VPR-RIR	0.720 ± 0.400	0.890 ± 0.043	0.690 ± 0.069	0.940 ± 0.042
	DEC	**0.390 ± 0.470**	**0.180 ± 0.036**	**0.430 ± 0.380**	**0.085 ± 0.039**
	CBF	0.510 ± 0.480	0.230 ± 0.035	0.620 ± 0.500	0.150 ± 0.046
$\Delta_{RT30}(\downarrow)$	VPR-RIR	0.005 ± 0.020	0.027 ± 0.016	**0.002 ± 0.002**	0.035 ± 0.014
	DEC	**0.003 ± 0.003**	**0.001 ± 0.001**	0.005 ± 0.005	0.001 ± 0.001
	CBF	0.011 ± 0.004	0.004 ± 0.006	0.008 ± 0.007	**0.001 ± 0.001**
$\Delta_{RT60}(\downarrow)$	VPR-RIR	0.011 ± 0.040	0.053 ± 0.032	**0.005 ± 0.003**	0.069 ± 0.028
	DEC	**0.006 ± 0.006**	**0.002 ± 0.001**	0.009 ± 0.010	0.002 ± 0.001
	CBF	0.022 ± 0.009	0.008 ± 0.012	0.016 ± 0.013	**0.002 ± 0.003**
MSE(↓)	VPR-RIR	7.2e-4 ± 1.7e-3	1.0e-3 ± 9.1e-4	5.5e-5 ± 3.9e-5	7.7e-5 ± 5.0e-5
	DEC	**3.1e-5 ± 1.2e-4**	**3.4e-5 ± 4.5e-5**	**6.8e-7 ± 5.0e-7**	**1.5e-6 ± 9.9e-7**
	CBF	2.7e-4 ± 1.0e-3	3.2e-4 ± 4.0e-4	2.5e-6 ± 1.8e-6	4.5e-6 ± 2.6e-6

Table 1: *Median ± standard deviation error scores for the experiments with additive Gaussian noise,* $SNR = 40$ *dB*

The reconstructed reverberant speech signal shows a significantly higher median PESQ score for VPR-RIR (around 4.0) compared to DEC (around 3.4) and CBF (around 3.3). Additionally, the interquartile range is narrower for VPR-RIR. As for the STOI score, VPR-RIR also exhibits a slightly better median, with a much narrower interquartile range compared to both DEC and CBF.

In conclusion, these results suggest that VPR-RIR , guided by a physical model and a variational approach, demonstrates greater stability in this setting and highlights its potential for future work in dereverberation tasks, especially when the dry signal is unknown.

	VPR-RIR	DEC	CBF
$\Delta_{C80}(\downarrow)$	**0.350 ± 0.190**	2.200 ± 2.900	3.400 ± 3.200
$\Delta_{D50}(\downarrow)$	**0.010 ± 0.013**	0.110 ± 0.100	0.160 ± 0.100
$\Delta_{DRR}(\downarrow)$	**0.000 ± 0.001**	0.002 ± 0.008	0.001 ± 0.009
$\Delta_{EDC}(\downarrow)$	**0.005 ± 0.013**	0.075 ± 0.049	0.090 ± 0.053
$\Delta_{EDR}(\downarrow)$	**0.770 ± 0.053**	1.000 ± 0.510	1.100 ± 0.560
$\Delta_{RT30}(\downarrow)$	**0.002 ± 0.001**	0.084 ± 0.045	0.092 ± 0.043
$\Delta_{RT60}(\downarrow)$	**0.005 ± 0.002**	0.170 ± 0.091	0.180 ± 0.086
MSE(↓)	**0.001 ± 0.001**	0.008 ± 0.028	0.010 ± 0.028

Table 2: *Median ± standard deviation error scores with additive noise from WHAMR! Noise,* $SNR = 20$ *dB*

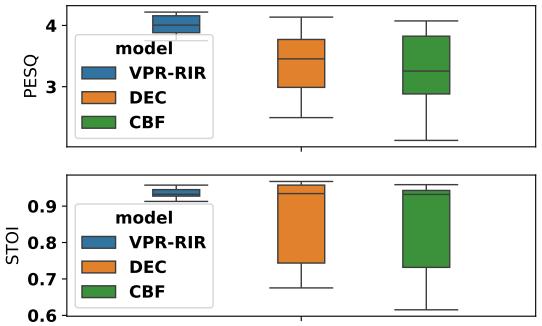

Figure 2: *Boxplot of the PESQ, STOI scores for the methods with additive noise from WHAMR! Noise* $SNR = 20$ *dB*

5. Conclusion & Future Works

In this paper, we propose a new method for estimating the Room Impulse Response (RIR), derived from an existing stochastic and physics-informed reverberation model presented in [21]. Our approach is based on a variational formulation, which not only allows for accurate RIR estimation but also provides the corresponding filter used to model the RIR. Results obtained from noisy reverberant signals demonstrate the effectiveness of our robust RIR estimator compared to traditional baselines.

In future work, this method can be extended to more challenging Acoustic Matching tasks, where a sound source is located in Room A, and the goal is to transfer it to Room B. The mixing process can be represented as a product of matrices, making it feasible to apply a variational approach to estimate the relevant parameters. Additionally, we plan to investigate convergence improvements, possibly through an expectation-maximization framework where parameters are updated alternately. Another direction for future work involves adapting the proposed method to the recently developed statistical wave field theory, which defines the statistical properties of wave propagation in bounded domains [32].

6. Acknowledgements

This work was supported by ANR Project SAROUMANE (ANR-22-CE23-0011).

7. References

[1] N. R. Shabtai, Y. Zigel, and B. Rafaely, "Room volume classification from room impulse response using statistical pattern recognition and feature selection," *The Journal of the Acoustical Society of America*, vol. 128, no. 3, pp. 1155–1162, 2010.

[2] J. Su, Z. Jin, and A. Finkelstein, "Acoustic matching by embedding impulse responses," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 426–430.

[3] S. Lee, H.-S. Choi, and K. Lee, "Yet another generative model for room impulse response estimation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2023, pp. 1–5.

[4] A. Tsilfidis, I. Mporas, J. Mourjopoulos, and N. Fakotakis, "Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing," *Computer Speech & Language*, vol. 27, no. 1, pp. 380–395, 2013.

[5] P. P. Parada, D. Sharma, P. A. Naylor, and T. v. Waterschoot, "Reverberant speech recognition exploiting clarity index estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, pp. 1–12, 2015.

[6] A. Ratnarajah, I. Ananthabhotla, V. K. Ithapu, P. Hoffmann, D. Manocha, and P. Calamia, "Towards improved room impulse response estimation for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[7] A. Belhomme, R. Badeau, Y. Grenier, and E. Humbert, "Amplitude and phase dereverberation of harmonic signals," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 294–298.

[8] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1717–1731, 2010.

[9] K. Sekiguchi, Y. Bando, A. A. Nugraha, M. Fontaine, K. Yoshii, and T. Kawahara, "Autoregressive moving average jointly-diagonalizable spatial covariance analysis for joint source separation and dereverberation," *IEEE/ACM Trans. ASLP*, vol. 30, pp. 2368–2382, 2022.

[10] L. Bahrman, M. Fontaine, J. Le Roux, and G. Richard, "Speech dereverberation constrained on room impulse response characteristics," in *Proc. INTERSPEECH*, 2024, pp. 622–626.

[11] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021.

[12] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE/ACM Trans. ASLP*, 2023.

[13] J. Su, Z. Jin, and A. Finkelstein, "Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," in *Proc. INTERSPEECH*, 2020.

[14] M. 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.

[15] J.-D. Polack, "La transmission de l'énergie sonore dans les salles (*The transmission of sound energy in rooms*)," Ph.D. dissertation, Le Mans, 1988.

[16] T. Komatsu, G. Peters, T. Matsui, I. Nevat, and K. Takeda, "Modeling room impulse response via composites of spatial-temporal Gaussian processes," in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1. AIP Publishing, 2013.

[17] M. Jälmby, F. Elvander, and T. Van Waterschoot, "Low-rank room impulse response estimation," *IEEE/ACM Trans. ASLP*, vol. 31, pp. 957–969, 2023.

[18] D. Sundström, F. Elvander, and A. Jakobsson, "Estimation of impulse responses for a moving source using optimal transport regularization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 921–925.

[19] S. Lee, H.-S. Choi, and K. Lee, "Differentiable artificial reverberation," *IEEE/ACM Trans. ASLP*, vol. 30, pp. 2541–2556, 2022.

[20] I. Martin, F. Pastor, F. Fuentes-Hurtado, J. A. Belloch, L. Azpicueta-Ruiz, V. Naranjo, and G. Piñero, "Predicting room impulse responses through encoder-decoder convolutional neural networks," in *Proc. IEEE Int. Workshop Mach. Learn. Signal. Process.*, 2023, pp. 1–6.

[21] R. Badeau, "Common mathematical framework for stochastic reverberation models," *J. Acoust. Soc. Am.*, vol. 145, no. 4, pp. 2733–2745, Apr. 2019.

[22] A. Aknin and R. Badeau, "Stochastic reverberation model with a frequency dependent attenuation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2021, pp. 351–355.

[23] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[24] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv e-prints*, pp. arXiv–1412, 2014.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 5206–5210.

[27] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černocký, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, Aug. 2019.

[28] C. Borß, "A VST Reverberation Effect Plugin Based on Synthetic Room Impulse Responses," -, 2009.

[29] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "WHAMR!: Noisy and Reverberant Single-Channel Speech Separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2020, pp. 696–700.

[30] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2001, pp. 749–752 vol.2.

[31] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4214–4217.

[32] R. Badeau, "Statistical wave field theory," *J. Acoust. Soc. Am.*, vol. 156, no. 1, pp. 573 – 599, Jul. 2024.