Sparse Parametric Modeling of the Early Part of Acoustic Impulse Responses

Constantinos Papayiannis, Christine Evers and Patrick A. Naylor
Department of Electrical and Electronic Engineering, Imperial College London, UK
{papayiannis, c.evers, p.naylor}@imperial.ac.uk

***Abstract*—Acoustic channels are typically described by their Acoustic Impulse Response (AIR) as a Moving Average (MA) process. Such AIRs are often considered in terms of their early and late parts, describing discrete reflections and the diffuse reverberation tail respectively. We propose an approach for constructing a sparse parametric model for the early part. The model aims at reducing the number of parameters needed to represent it and subsequently reconstruct from the representation the MA coefficients that describe it. It consists of a representation of the reflections arriving at the receiver as delayed copies of an excitation signal. The Time-Of-Arrivals of reflections are not restricted to integer sample instances and a dynamically estimated model for the excitation sound is used. We also present a corresponding parameter estimation method, which is based on regularized-regression and nonlinear optimization. The proposed method also serves as an analysis tool, since estimated parameters can be used for the estimation of room geometry, the mixing time and other channel properties. Experiments involving simulated and measured AIRs are presented, in which the AIR coefficient reconstruction-error energy does not exceed 11.4% of the energy of the original AIR coefficients. The results also indicate dimensionality reduction figures exceeding 90% when compared to a MA process representation.**

***Index Terms*—Sparse Modeling; Reverberation; Acoustic Environments; Reflection TOA Estimation.**

I. Introduction

Reverberation is observed in almost all real-life acoustic environments as sound emitted by a source is reflected off the boundaries of the enclosure and the surfaces of objects. It offers warmth to the sound and it is desirable in the case of music [1]. In the case of speech, it can degrade the performance of Automatic Speech Recognition (ASR) systems and impact intelligibility [2]. Reverberation is the focus of many applications which aim at either recreating it or combating its negative effects. The reverberation effect of a stationary acoustic environment can be described by the system's Acoustic Impulse Response (AIR). AIRs consist of three parts, the direct path sound, the early and the late reflections [2]. The early part refers to the extent of the AIR arising from strong, discrete reflections covering a duration up to the so-called mixing time. The effect of the early reflections as perceived by the listener is a change in the timbre of the original sound, with the effect referred to as coloration. The perceptual effect of the late reflections is a prolonging of the original sound.

AIRs are typically modeled as a Moving Average (MA) process which involves thousands of coefficients. However, in many applications low-dimensionality is needed due to constraints in memory and computational power. Such applications involve dereverberation [3], auralization [4] and environment classification [5]. Low-dimensionality also benefits tasks involving an estimation of the channel as low-order models are often easier to estimate robustly. Alternative representations to the MA process representation have been previously proposed in the literature. Early examples involve Autoregressive Moving Average (ARMA) modeling [6]. The main motivation for their use is their ability to model room-modes, an important aspect of room acoustics [7]. However, issues such as model order selection and the increase in room-mode density beyond the Schroeder frequency [7] limit their practical use. More modern approaches include the use of Kautz filters [8] and a direct search for the salient characteristics of acoustic environments from the AIRs using dimensionality reduction methods [9]. Existing approaches for the targeted modeling and simulation of the late reflections [10], [11] are based on their stochastic nature, for which a statistical description is more appropriate. Early reflections however follow a structured distribution in time, related to room geometry and source and receiver positions.

We propose a novel approach for the construction of a parametric model for the early part of AIRs. Starting from the MA model coefficients, the number of parameters required for its representation is reduced. It is also aimed at reconstructing the coefficients, using the resulting representation. The model describes the process of sampling a sound field which is composed of sound rays propagating in the enclosure and being reflected off its boundaries and the surface of objects. It incorporates a dynamic representation for the excitation signal which is emitted by the source and a set of reflections modeled as superimposed delayed copies of it. A two stage optimization method is employed for the parameter estimation. In the first step, an initial set of values is obtained by approximating the problem using linear regression followed by fine-tuning of the parameter values by optimization in local time-regions.

The proposed approach offers a variety of advantages over previously proposed methods. The reduction in dimensionality it offers is attributed to the exploitation of the structure of the AIR, which is characterized by the sparse nature of strong early acoustic reflections. It is not attributed to an approximation of the channel by under-fitting the underlying model. During the model fitting process, reflections are detected from the AIR coefficients, which are subsequently described by their amplitudes and Times-of-Arrival (TOAs). The proposed

ISBN 978-0-9928626-7-1 © EURASIP 2017

method, during this process, does not bound the TOAs to integer sample instances, it models the excitation sound and makes no assumptions about the number of overlapping reflections. These are some of the disadvantages exhibited by existing methods [12], [13], [14].

The structure of the remainder of this paper is as follows: Section II formulates the model and proposed parameter estimation method, Section III describes the experimental work done and a conclusion is given in Section IV.

II. Method

A. Signal Model

AIRs are measured by exciting an acoustic environment with an excitation signal $h_e(n)$. A measured AIR can be modeled as

$$h(n) = \sum_{i=1}^{D} \left[\beta_i h_e(n) * \frac{\sin\ \pi(n-k_i)}{\pi(n-k_i)} \right] + \nu(n) \tag{1}$$

for $n \in \{0, \ldots, N-1\}$, $*$ indicating a convolution process and $\nu(n)$ additive noise. The TOAs of reflections are represented by $k_i \in [0, \infty) \forall i \in \{1, \ldots, D\}$. When the TOA of a reflection is a sampling instance, the reflection contributes to the AIR as a delayed copy of $h_e(n)$, scaled by β_i. When this is not the case and under ideal band-limiting, in addition to the delay and scaling, $h_e(n)$ is convolved with the sinc function.

B. Initial Parameter Estimation

Without prior knowledge about any of the unknowns of (1), fitting the model to a measured AIR leads to a high-dimensional and nonlinear problem. A simplification to (1) can be made by temporarily assuming that the additive noise is negligible, followed by reformulating the problem as

$$\hat{h}_R(n) = \sum_{r=1}^{M} w_r x_r(n) = \mathbf{w}^T \mathbf{x}(n) \tag{2}$$

$$x_r(n) = h_e(n) * \frac{\sin\ \pi\left(n - \frac{r-1}{Q}\right)}{\pi\left(n - \frac{r-1}{Q}\right)}, \tag{3}$$

where $\mathbf{w} = [w_1, \ldots, w_M]^T$ and $\mathbf{x}(n) = [x_1(n), \ldots, x_M(n)]^T$. M is an integer multiple of the number of coefficients of the measured AIR, N. The integer $Q = {}^{M}/_{N}$ defines the number of candidate TOAs of reflections per AIR coefficient. This reformulation simplifies the problem by bounding and quantizing the space of possible TOAs and transforming it to a linear-regression form [15]. All values in the space are considered as candidate TOAs of reflections and the only remaining unknowns are their amplitudes $\mathbf{w}$. The unknown number of reflections is estimated as $\hat{D} = \|\mathbf{w}\|_0$, where $\|\cdot\|_0$ counts the number of non-zero elements of the vector. These elements correspond to the scaling coefficients β_i in (1). For the early part of the AIR, $\mathbf{w}$ is expected to be sparse as the reflection density is low.

A Least Squares (LS) solution is expected to yield many non-zero values in $\mathbf{w}$ [15] and lead to overestimates of D. One way to deal with this problem is regularization. The Least Absolute Shrinkage and Selection Operator (LASSO) [16] is appropriate for the task as it promotes sparsity. Minimizing the residual error between $\hat{h}_R(n)$ and $h(n)$ however will give emphasis to describing low-order reflections which have the highest amplitude. In order to account for this, $\tilde{h}(n)$ is created which represents the Energy Decay Curve (EDC) compensated AIR [14]. LASSO is then used to find $\tilde{\mathbf{w}}$ that minimizes the expression

$$e = \left\| \tilde{h}(n) - \hat{h}_R(n) \right\|_2 + \lambda \left\| \tilde{\mathbf{w}} \right\|_1, \tag{4}$$

where λ is a scalar constant. The solution for $\tilde{\mathbf{w}}$ corresponds to the EDC compensated AIR [14]. In order find the vector $\mathbf{w}$ which corresponds to the original AIR, the dot product of $\tilde{\mathbf{w}}$ and the square-root of the EDC is taken.

C. Model Parameter Fitting

Despite the regularization imposed in (4), more than one nonzero adjacent coefficients in $\mathbf{w}$ might correspond to the same reflection. This can be attributed to the quantization of the time domain which was done in (2). Regularization might also lead to underestimates of the reflection amplitudes as it directly penalizes them. In order to more accurately estimate the parameter values, further optimization is performed with $\mathbf{w}$ used for its initialization.

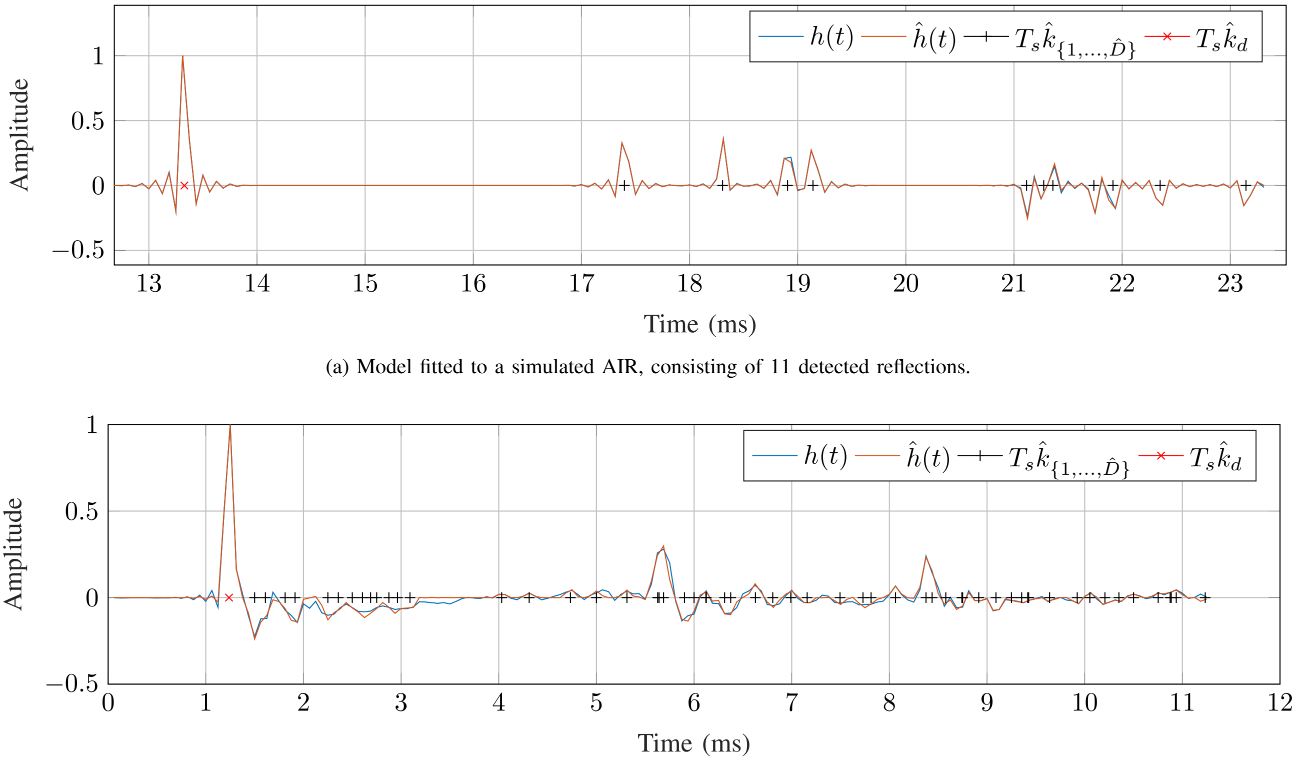
The interior-point method [17] is used to optimize parameter values by minimizing the Mean Square Error (MSE) between the final model and windows of the AIR. At the beginning of the process we set $\hat{D} = 0$, $\hat{\boldsymbol{\beta}} = \emptyset$ and $\hat{\boldsymbol{k}} = \emptyset$. Each optimization process considers Q elements of $\mathbf{w}$, defined as $\mathbf{w}_0$. The initialization for the amplitude and TOA optimization of a number of possible reflections is then given by

$$\boldsymbol{\beta}_0 = \mathbf{w}_0 \cap \mathbb{R}_{\neq 0} \tag{5}$$

$$\boldsymbol{k}_0 = \left\{ \frac{r-1}{Q}\ :\ w_{0,r} \neq 0 \right\}. \tag{6}$$

The TOAs are bounded within the range of TOAs described by $\mathbf{w}_0$. Each AIR window used to evaluate the MSE extends a time τ_e at each side of this range. Ideally, the entire AIR could be used for the evaluation of the MSE, however practical limitations prohibit this. Higher τ_e values lead to a decrease in the residual-error but increased computation times. To promote sparsity, the number of possible reflections tested spans the range $\{0, \ldots, \|\mathbf{w}_0\|_0\}$. This provides a MSE value for each case. The model with the maximum number of reflections that provides a decrease of s to the MSE with regards to the previous one is accepted, where s is a scalar. The accepted number of reflections is added to $\hat{D}$ and the parameters are appended to the vectors $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{k}}$, before moving to the next AIR window.

Setting $s < 1$ favours the choice of models with lower dimensionality by expecting the addition of more parameters to significantly reduce the MSE. Sparsity is also promoted indirectly as multiple elements of $\mathbf{w}_0$ that may be targeting a



(a) Model fitted to a simulated AIR, consisting of 11 detected reflections.

(b) Model fitted to an AIR measured in a lecture room, consisting of 57 detected reflections.

Fig. 1: Results of model fitting to simulated and recorded AIRs sampled at 16 kHz, T_s indicates the sampling period.

single reflection, are at this point "fused" to a single parameter-pair, with the same TOA and amplitude. Repeating the process for all windows, results in a model for the AIR given by

$$\hat{h}(n) = \sum_{i=1}^{\hat{D}} \hat{\beta}_i \hat{h}_e(n) * \frac{\sin\ \pi(n-\hat{k}_i)}{\pi(n-\hat{k}_i)}. \tag{7}$$

D. Excitation Model

Nonidealities of the AIR being modeled include factors such as band-limiting at the source and receiver. These are accounted for by considering equivalent nonidealities in a corresponding excitation for the AIR. In related work, this excitation was modeled as a modulated Gaussian pulse [18]. This model is adopted here, leading to the following expression for the excitation signal

$$\hat{h}_e(n) = e^{-\theta^2(nT_s)^2}\cos(2\pi f_e n T_s), \tag{8}$$

$$h_d(n) = \hat{\beta}_d \hat{h}_e(n) * \frac{\sin\ \pi(n-k_d)}{\pi(n-k_d)}, \tag{9}$$

where k_d is the TOA of the direct sound and T_s is the sampling period. We estimate the parameters θ and f_e from window $h_d(n)$ of the AIR containing the direct sound. Assuming that the direct sound will have the highest energy, this window of length 4 ms is centered around the highest energy sample. Estimates for the parameters are then found using [19], which aims to find the global minimum of the MSE between $\hat{h}_e(n)$ and $h_d(n)$. For each tested parameter set, k_d is varied to best align $\hat{h}_e(n)$ to $h_d(n)$ and $\hat{\beta}_d$ is set to match the maximum values between the two. The benefit of creating a model to represent the excitation signal over using samples directly extracted from the AIR is avoiding the inclusion of any reflections or noise in the representation.

E. Adjusting Regularization

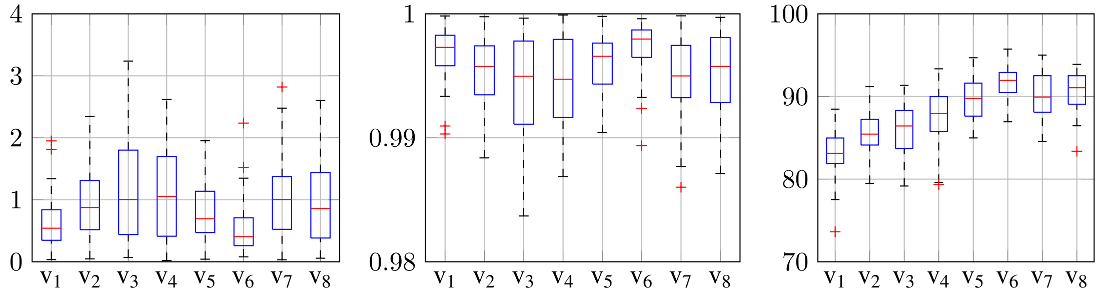
For each AIR the value of λ for (4) is self-adjusted by first finding λ_0, the first value for which $\|\mathbf{w}\|_0 > 0$. Subsequently, LASSO is run for $\lambda_\psi = \lambda_0 \cdot 10^{-0.04\psi}$, where $\psi \in \{1,\dots,100\}$. Based on the MSE values ϵ_ψ, we find the λ_m value which leads to the minimum MSE value ϵ_m. The final λ value is chosen based on the largest ψ index for which $\epsilon_\psi < \epsilon_m + \frac{\sigma_\epsilon}{3}$. This provides a trade-off between sparsity and accuracy [20].

III. Experiments

The results of fitting the model to simulated and measured AIRs are illustrated in the following experiments. All AIRs considered are sampled at 16 kHz.

Fig. 1a shows the result for an AIR simulated using [21] for a "shoe-box" room. The room dimensions were [3.5, 3.9, 2.5] m, with the last dimension indicating the height. The source and receiver were placed in a random location within the room, drawn from a uniform distribution, at least 1 m from the room's boundaries. For visual clarity, the results are shown for the first 10 ms after the arrival of the direct sound at the receiver. The result of the model fitting process (see Fig. 1a)

ISBN 978-0-9928626-7-1 © EURASIP 2017

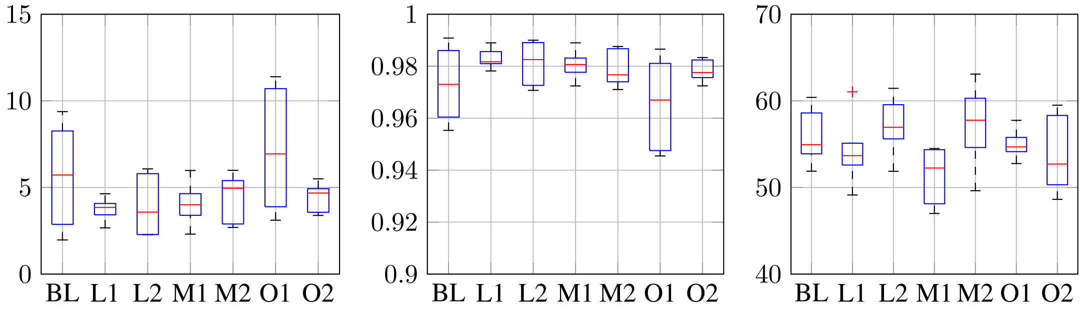


(a) Normalized Error (%) by room-volume group

(b) Correlation by room-volume group

(c) Dimensionality Reduction (%) by room-volume group

Fig. 2: Results for 480 simulated AIRs in 15 rooms, grouped with increasing indices indicating a larger room volume.
Group Legend: v1: 15-24 m^3, v2: 25-34 m^3, v3: 35-44 m^3, v4: 55-64 m^3, v5: 65-74 m^3, v6: 85-94 m^3, v7: 95-104 m^3, v8: 105-114 m^3



(a) Normalized Error (%) by room

(b) Correlation by room

(c) Dimensionality Reduction (%) by room

Fig. 3: Results for 42 measured AIRs in 7 rooms, AIRs are grouped in terms of rooms.
Room Legend : BL:Building Lobby, L1: Lecture Room 1, L2: Lecture Room 2, M1: Meeting Room 1, M2: Meeting Room 2, O1: Office 1, O2: Office
Room Volume: BL: 72 m^3, L1: 200 m^3, L2: 360 m^3, M1: 92 m^3, M2: 250 m^3, O1: 47 m^3, O2: 48 m^3

appears to almost perfectly model the AIR when comparing $\hat{h}(t)$ to $h(t)$.

The same process was repeated for a measured AIR, part of the Acoustic Characterization of Environments (ACE) Database [22]. The recording took place in a lecture room with dimensions [6.9, 9.7, 3.0] m and the receiver is one of the channels of a 32 microphone array. Similarly to the previous case, by comparing $\hat{h}(t)$ to $h(t)$ (see Fig. 1b), we can see that the model captures the structure of the AIR. The model for the excitation signal also accurately represents the direct sound at the indicated position.

To gain objective insight into the performance of the model, the normalized error, Pearson's correlation coefficient and dimensionality reduction values were extracted from the results. These are respectively denoted by ψ, ρ and ζ. The normalized error is defined as

$$\psi = \frac{\|h(n) - \hat{h}(n)\|_2}{\|h(n)\|_2} \times 100\%, \tag{10}$$

which expresses the residual error as a percentage of the overall AIR energy. The dimensionality reduction is evaluated as the percentage

$$\zeta = \left(1 - \frac{2\hat{D} + 4}{N}\right) \times 100\%, \tag{11}$$

where N the number of taps required for a MA process representation. In these examples $N = k_d + (10 \cdot 10^{-3}\ \mathrm{s}) \cdot (16 \cdot 10^3\ \mathrm{Hz})$ as the first 10 ms after the arrival of the direct path are considered. The number of parameters used by the model involve the TOAs and amplitudes of $\hat{D}$ reflections and four parameters to describe the direct excitation model of (8). In terms of Fig. 1a, $\psi = 0.28\%$, $\rho = 0.999$ and $\zeta = 96\%$. In terms of Fig. 1b, $\psi = 2.61\%$, $\rho = 0.987$ and $\zeta = 67\%$. For the model fitting process (see Sec. II-C), s and τ_e were set to 0.90 and 0.5 ms respectively. This provided sparse solutions and a trade-off between residual error and computation times.

To further illustrate the performance of the proposed modeling approach two larger experiments are presented. In the first one, 480 AIRs were simulated. This involved 15 rooms of different dimensions with a single source and 32 receivers. Source and receivers were randomly placed, at least 1 m away from the boundaries. The model was fitted to the first 24 ms after the arrival of the direct sound for each. This is defined as the mixing time in [23]. The results are shown in Fig. 2. Rooms have been split into 8 groups based on their volume. The normalized error in Fig. 2a and correlation in Fig. 2b highlight that the modeling accuracy remains constant across all room volume groups. The dimensionality reduction results in Fig. 2c show that more parameters are required to model smaller rooms. This increase in dimensionality for decreasing room sizes is due to the more rapidly increasing reflection density in smaller enclosures.

For the second experiment the task was to model the 42 AIRs recorded using a 3 microphone mobile phone array in 7 rooms provided in the ACE database [22]. This corresponds to 2 sets of measurements, with the receiver positions varying between the two and the rest of the setup unchanged. In this case, results are grouped by individual rooms instead of room volume and are sorted by room type. Furthermore, higher indices for a specific room type indicate a higher room volume, i.e. Lecture Room 2 has a higher volume than Lecture Room 1 and so on. The higher error in Fig. 3a and lower correlation values in Fig. 3b for this experiment indicate the increased challenges in modeling measured AIRs. The dimensionality reduction is also lower overall, as shown in Fig. 3c. For measured AIRs, the model has to include further parameters to account for reflections from objects other than the enclosure's boundaries. For instance, reflections from microphone-stands involved in the setup are expected to be present in measured data, which is not the case for simulations. Furthermore, ambient and sensor noise in AIR recordings is expected to impact the residual error.

In contrast to the first experiment which involved simulated AIRs, larger real rooms do not consistently lead to a higher dimensionality reduction. Nevertheless, higher reductions in dimensionality are still shown between rooms of the same type as their volume increases. This indicates that volume is again a factor to be considered in terms of the model's dimensionality. The normalized error in Fig. 3a shows high variability for specific rooms across AIRs. The two rooms with the highest variability are the Building Lobby and Office 1. Investigating further indicates that the measurement position was highly correlated to the level of error. For both rooms, the AIRs with the highest modeling error involved the receiver being placed closest to a room wall. The opposite was true for receiver locations closest to the middle of the room. AIRs closer to the room boundaries therefore appear to be more challenging to model. This is attributed to the fact that reflection spacing will be smaller, making the modeling task more difficult.

Another remark which further illustrates the usefulness of the proposed method is its ability to estimate the TOAs of reflections in AIRs, which makes it a valuable analysis tool. This is illustrated by Figs. 1a and 1b. Estimated TOAs can be used to estimate room geometry [12], the mixing time [14] and other channel properties.

IV. Conclusion

A novel approach for the construction of a low dimensional parametric model of the early part of AIRs has been presented. Experiments involving simulated and measured AIRs, showed that using the proposed modeling approach can reduce the dimensionality by more than 90% and 60% respectively. The corresponding AIR coefficient reconstruction normalized-error does not exceed 3.2% for simulated and 11.4% for measured AIRs.

References

[1] M. Ermann, *Architectural Acoustics Illustrated*, John Wiley & Sons, 2015.

[2] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer-Verlag, 2010.

[3] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 45–48.

[4] P. Samarasinghe, T. Abhayapala, M. Poletti, and T. Betlehem, "An Efficient Parameterization of the Room Transfer Function," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 2217–2227, 2015.

[5] C. Papayiannis, C. Evers, and P. A. Naylor, "Discriminative Feature Domains for Reverberant Acoustic Environments," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, Louisiana, USA, Mar. 2017, (accepted).

[6] Y. Haneda, S. Makino, Y. Kaneda, and N. Koizumi, "ARMA modeling of a room transfer function at low frequencies," *J. Audio Eng. Soc. of Japan*, vol. 15, pp. 353–355, Sept. 1994.

[7] H. Kuttruff, *Room Acoustics, Fifth Edition*, CRC Press, 2009.

[8] 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 *Proc. European Signal Processing Conf. (EUSIPCO)*, Sept. 2014, pp. 1–5.

[9] R. Duraiswami and V. C. Raykar, "The Manifolds of Spatial Hearing," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, Pennsylvania, USA, Mar. 2005, pp. 285–288.

[10] E. Lehmann and A. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1429–1439, Aug. 2010.

[11] P. Rubak and L. G. Johansen, "Artificial reverberation based on a pseudo-random impulse response II," in *Proc. Audio Eng. Soc. (AES) Convention*, May 1999.

[12] F. Antonacci, J. Filos, M. Thomas, E. Habets, A. Sarti, P. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2683–2695, Dec. 2012.

[13] I. Kelly and F. Boland, "Detecting arrivals in room impulse responses with dynamic time warping," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 7, pp. 1139–1147, July 2014.

[14] G. Defrance, L. Daudet, and J. D. Polack, "Using matching pursuit for estimating mixing time within room impulse responses," *Acta Acustica united with Acustica*, vol. 95, no. 6, pp. 1071–1081, 2009.

[15] D. Ba, F. Ribeiro, C. Zhang, and D. Florencio, "L1 regularized room modeling with compact microphone arrays," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA, Mar. 2010, pp. 157–160.

[16] S. Theodoridis, *Machine Learning*, Academic Press, 2015.

[17] W. Forst and D. Hoffmann, *Optimization—Theory and Practice*, Springer-Verlag, 2010.

[18] C. H. Jeong, J. Brunskog, and F. Jacobsen, "Room acoustic transition time based on reflection overlap," *J. Acoust. Soc. Am.*, vol. 127, pp. 2733–2736, 2010.

[19] Z. Ugray, L. Lasdon, J. Plummer, F. Glover, J. Kelly, and R. Martí, "Scatter Search and Local NLP Solvers: A Multistart Framework for Global Optimization," *INFORMS Journal on Computing*, vol. 19, no. 3, pp. 328–340, July 2007.

[20] L. Breiman, J. H. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, CRC Press, 1984.

[21] C. J. A. Wabnitz, N. Epain and A. Schaik, "Room acoustics simulation for multichannel microphone arrays," in *Proc. Intl. Symp. on Room Acoustics (ISRA)*, Melbourne, Australia, Aug. 2010.

[22] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 10, pp. 1681–1693, Oct. 2016.

[23] J.-D. Polack, *La Transmission de l'énergie Sonore Dans Les Salles*, Ph.D. thesis, Université du Maine, Le Mans, France, Dec. 1988.