KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Antonello N., De Sena E., Moonen M., Naylor P. A., van Waterschoot T. (2017) **Room impulse response interpolation using a sparse spatio-temporal representation of the sound field** IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 25, no. 10, pp. 1929-1941, 2017.
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	
Journal homepage	http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=6570655
Author contact	niccolo.antonello@esat.kuleuven.be *+ 32 (0)16 321855*
IR	

(article begins on next page)

Room impulse response interpolation using a sparse spatio-temporal representation of the sound field

Niccolò Antonello, Enzo De Sena, *Member, IEEE,* Marc Moonen, *Fellow, IEEE,* Patrick A. Naylor, *Senior Member, IEEE,* and Toon van Waterschoot *Member, IEEE*

***Abstract*—Room Impulse Responses (RIRs) are typically measured using a set of microphones and a loudspeaker. When RIRs spanning a large volume are needed, many microphone measurements must be used to spatially sample the sound field. In order to reduce the number of microphone measurements, RIRs can be spatially interpolated. In the present study, RIR interpolation is formulated as an inverse problem. This inverse problem relies on a particular acoustic model capable of representing the measurements. Two different acoustic models are compared: the plane wave decomposition model and a novel time-domain model that consists of a collection of equivalent sources creating spherical waves. These acoustic models can both approximate any reverberant sound field created by a far field sound source. In order to produce an accurate RIR interpolation, sparsity regularization is employed when solving the inverse problem. In particular, by combining different acoustic models with different sparsity promoting regularizations, spatial sparsity, spatio-spectral sparsity and spatio-temporal sparsity are compared. The inverse problem is solved using a matrix-free large scale optimization algorithm. Simulations show that the best RIR interpolation is obtained when combining the novel time-domain acoustic model with the spatio-temporal sparsity regularization, outperforming the results of the plane wave decomposition model even when far fewer microphone measurements are available.**

***Index Terms*—Room impulse response, Microphone array, Sparse sensing, Inverse problems, Large scale optimization**

I. Introduction

Room Impulse Responses (RIRs) play a fundamental role in room acoustics. They not only provide useful parameters that describe sound fields qualitatively, such as reverberation time or clarity, but also their knowledge is fundamental in many applications such as channel equalization and sound field reproduction. A RIR between two points in a room is typically measured by generating a deterministic signal, e.g. a sine sweep, using a loudspeaker at one point and recording the sound pressure with a microphone at the other point [1], [2]. Measuring RIRs over a large volume can be a tedious and time-consuming task unless many microphones or a moving microphone [3] are available. Typically, one has to repeat the measurement multiple times for all the source and microphone positions of interest. In order to avoid this, an alternative lies in the spatial interpolation of the measured RIRs to obtain estimates of the RIRs at positions where no microphone measurements are available. In [4] the space-time spectrum of the *plenacoustic function* is studied. In practice, the plenacoustic function represents the collection of all the RIRs associated with a room as a function of space and time. It is shown that the plenacoustic function space-time spectrum is band limited, implying that the Nyquist sampling theorem can be applied. This spatio-temporal interpolation however still requires many microphone measurements. For example, if a 3-dimensional uniform microphone array is used, the RIRs between the microphones can be interpolated if the microphone measurements are spaced by the distance

$$X < \frac{c}{2F_u}, \tag{1}$$

where F_u is the cut-off frequency in Hz of the sound source generating the sound field. Here c represents the speed of sound for which in this paper 343 m/s is used.

In general, it is necessary to seek a milder criterion than that given by the Nyquist sampling theorem. Compressed Sensing (CS) represents such an alternative [5]: this framework consists of solving *optimization problems* where the lack of information due to a limited number of available measurements is compensated for by exploiting *prior knowledge*. In particular, in CS this prior knowledge consists of the fact that the sought solution is *sparse*. In this context, CS represents a particular case of an *inverse problem*. As a matter of fact, inverse problems also seek for variables that are not directly measurable [6]. Inverse problems rely on describing the physical phenomenon under study using a model that is partially unknown. These problems are in general *ill-posed*, which implies that meaningful solutions can only be obtained when the inverse problem is *regularized* using a specific prior knowledge. As in CS, inverse problems are posed as optimization problems and the CS framework in fact represents an inverse problem posed with a sparsity promoting regularization.

RIR interpolation has been posed in a wide variety of fashions: different acoustic models have been used, each one having its own specific parameters to estimate, and different

N. Antonello, M. Moonen, T. van Waterschoot are with KU Leuven, ESAT–STADIUS, Stadius Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, 3001 Leuven, Belgium.

Toon van Waterschoot is also with KU Leuven, ESAT–ETC, Advise Lab, Kleinhoefstraat 4, 2440 Geel, Belgium.

E. de Sena is with Institute of Sound Recording University of Surrey Guilford, Surrey, GU2 7XH, UK

P. A. Naylor is with Electrical & Electronic Engineering Imperial College Exhibition Road London SW7 2AZ, UK

This research work was carried out at the ESAT Laboratory of KU Leuven, in the frame of (i) the FP7-PEOPLE Marie Curie Initial Training Network "Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS)", funded by the European Commission under Grant Agreement no. 316969, (ii) KU Leuven Research Council CoE PFV/10/002 (OPTEC), (iii) KU Leuven Internal Funds C2-16-00449 'Distributed Digital Signal Processing for Ad-hoc Wireless Local Area Audio Networking', (iv) KU Leuven Impulsfonds IMP/14/037. The scientific responsibility is assumed by its authors.

algorithms have been used to solve the underlying optimization problems. For example, in [7]–[9] RIR interpolation is achieved by solving a classical inverse problem where the acoustic impedances of the walls are the parameters to be estimated. It is assumed that the geometry of the room is known and the wave equation is discretized using wave-based numerical methods such as the Finite Difference Time-Domain (FDTD) method [8], [9] or the Boundary Element Method (BEM) [7]. The acoustic impedance of the walls is reconstructed by solving a non-convex optimization problem. Once the acoustic properties of the walls are determined, the RIRs can be generated at every position of the room using the underlying wave-based numerical method. Nevertheless, these wave-based numerical methods suffer from numerical errors and are only accurate at low frequencies [10]. Moreover, they require a precise knowledge of the room geometry and lead to non-convex optimization problems.

Other approaches that do not require the knowledge of the room geometry have been proposed. For example many approaches rely on a spatial parametrization of the sound field, obtained as well from physical models, i.e. from the spherical harmonic solutions of the wave equation [11]–[14]. The main idea is to extrapolate a finite set of parameters out of the measured RIRs such that the parametrization allows the sound field to be predicted at positions where no microphone measurement was made. Such a parametrization is strongly linked to another widely used acoustic model, the Plane Wave Decomposition Method (PWDM). In [15] it is shown that, as in the case of spherical harmonics, an acoustic sound field can be correctly approximated using a finite number of *plane waves* independently of the boundaries. Based on this theoretical result, many authors have chosen the PWDM as their acoustic model. For example, in [16] the room modes are identified using common acoustical poles techniques and then reconstructed using the PWDM using a fixed number of plane waves. Still in [16], alternatively a limited number of damped plane waves is chosen out of a large dictionary of damped plane waves using a greedy optimization algorithm. A similar technique is used in [17], where for each frequency an optimization problem is solved in order to find a sparse set of plane waves from a large dictionary that can interpolate the RIRs needed for multi-zone sound field reproduction. A different acoustic model is used in [18] where a wide-band RIR interpolation of the early part of the RIRs is performed. Here the sound field is modeled using the Image Method (IM) [19]. This inverse problem can be seen as to a localization problem: if the location of the image sources can be found, the sound field can be reconstructed. Inverse problems appear also in the acoustic holography field. Here yet another acoustic model of importance is the Equivalent Source Method (ESM), which consists of a collection of equivalent sources generating spherical waves [20]–[22].

These techniques and acoustic models are very much related to those used in sound field reproduction, for example Wave Field Synthesis (WFS) [23], the Spatial Decomposition Method (SDM) [24], and Pressure Matching (PM) [25], [26] where sparsity promoting regularizations have been used as well [27], [28]. While the aim of these methods is to reproduce a sound field, RIR interpolation techniques seek to estimate the sound field and require different treatments and formulations.

In many of these contexts, inverse problems lead to large scale optimization problems hence requiring specific optimization algorithms. For example, in [18] and [16], the optimization problems are solved using greedy algorithms, specifically with a modified version of Matching Pursuit (MP). An alternative way of solving large scale problems relies on first-order optimization algorithms, where matrix inversions are avoided. The most well-known first-order optimization algorithm is the gradient descent algorithm, which exhibits a low convergence rate and is not well suited for non-smooth cost functions that typically appear with sparsity promoting regularizations. Recent interest in large scale optimization problems has led to the development of accelerated first-order optimization algorithms which enjoy faster convergence and can deal with non-smooth cost functions. A successful family of first-order algorithms is referred to as Forward-Backward Splitting (FBS) also known as the proximal gradient method [29]. These algorithms have recently been substantially accelerated using quasi-Newton methods [30], [31]. An important aspect of large scale problems is the memory requirement which can easily become prohibitive. Matrix-free optimization avoids the storage of large matrices through the usage of the adjoint operators [32], [33].

In this paper a novel method for RIR interpolation is proposed. A time-domain acoustic model called Time-domain Equivalent Source Method (TESM) is described, which is a modified time-domain version of the ESM. This acoustic model is used to approximate the sound field in a source-free volume. The main differences with respect to ESM are that the source signal can be included in the model to improve RIR interpolation and that the formulation is in the time-domain, which allows *spatio-temporal sparsity* regularization to be applied. Due to the large number of equivalent sources the optimization problem becomes large-scale and the computational aspect becomes relevant. In particular matrix-free optimization is used in order to reduce the memory requirement and the algorithm proposed in [31] is used to speed up the convergence. The spatio-temporal sparsity regularization is compared with other regularizations: it is shown that when using the PWDM in the inverse problem, it is possible to impose either spatial sparsity or spatio-spectral sparsity, with the former being a better choice. Still, numerical simulations show that the choice of promoting spatio-temporal sparsity with TESM always outperforms the RIR interpolation obtained with PWDM and leads to a significant decrease in the number of microphone measurements required. In this study RIR interpolation is limited to the far-field case.

The paper is organized as follows: in Sections II and III the PWDM and the novel TESM are presented respectively. In Section IV, these acoustic models are used in the inverse problems that rely on different regularizations involving a sparse representation: spatial sparsity, spatio-spectral sparsity and spatio-temporal sparsity. In Section V the optimization algorithm used to solve these inverse problems is briefly described. In Section VI the efficient computation of the derivatives needed to solve the optimization problems is

given. Here an interpretation of the adjoint operators of the acoustic models needed in this computation is also given. In Section VII the simulation results are presented: firstly in VII-A the tuning of the level of the regularization is described, secondly in VII-B the RIR interpolation performances are compared for different acoustic models and regularizations. In Section VII-C the solutions of the inverse problems are visualized. In Section VIII experimental results are presented. Finally, conclusion in Section IX are given.

II. Plane wave decomposition method

In this section, the Plane Wave Decomposition Method (PWDM) is briefly reviewed. Plane waves are defined as

$$\hat{\phi}_{f,l}(\mathbf{x}) = e^{i\mathbf{k}_{f,l}^T\mathbf{x}} \tag{2}$$

and they are solutions of the homogeneous Helmholtz equation, that is

$$\nabla^2\hat{p}_f(\mathbf{x}) + k_f^2\hat{p}_f(\mathbf{x}) = 0 \text{ on } \mathbb{R}^3, \tag{3}$$

which assumes a time-harmonic behavior of the acoustic field, i.e. $p(\mathbf{x},t) = \mathfrak{Re}(\hat{p}_f(\mathbf{x})e^{i\omega_f t})$ for $\omega_f = 2\pi f F_s/N_f$. Here f is the frequency index, $\hat{p}_f(\mathbf{x}) : \mathbb{R}^3 \to \mathbb{C}$ is the complex sound pressure at a particular frequency f and ∇^2 is the Laplacian operator over the spatial variables $\mathbf{x}$. The wave number k_f is defined as $k_f = \omega_f/c$ and $\mathbf{k}_{f,l} = k_f\mathbf{n}_l \in \mathbb{R}^3$ is the wave vector, with $\mathbf{n}_l$ a unit vector that defines the direction of the lth plane wave. Equation (3) can be written for $f = 0,\dots,N_f-1$ discrete frequencies. In the following $\hat{\phi}_{f,l,m}$ will indicate $\hat{\phi}_{f,l}(\mathbf{x})|_{\mathbf{x}=\mathbf{x}_m}$ where $\mathbf{x}_m \in \mathbb{R}^3$ is a point in space.

A sound field in a source-free spatial domain $\Omega \subset \mathbb{R}^3$ can be well represented by a finite weighted sum of plane waves coming from N_w different directions [12]:

$$\hat{p}_f(\mathbf{x})|_{\mathbf{x}=\mathbf{x}_m} \approx \sum_{l=0}^{N_w-1} \hat{\phi}_{f,l,m}\hat{w}_{f,l} \text{ for } \mathbf{x}_m \in \Omega, \tag{4}$$

where the *weight* $\hat{w}_{f,l}$ is a complex scalar that weights the (f,l)-th plane wave. Under the assumptions that Ω is source-free, that the source generating the sound field and that the reflecting surfaces are in the far field, this acoustic model, known as the PWDM, gives a good approximation of any sound field [12], [15]. If these conditions are met, this approximation holds independently of the boundary conditions, the room geometry and the sound source type that generates the field outside Ω. Suppose that the weights $\hat{w}_{f,l}$ are available, if one wants to predict the sound pressure at N_m discrete positions $\mathbf{x}_m \in \Omega$ for $m = 0,\dots,N_m-1$, (4) can be generalized for discrete spatial points by

$$\hat{\mathbf{P}} = D_{pw}(\hat{\mathbf{W}}), \tag{5}$$

where $\hat{\mathbf{P}}$ is a matrix containing the complex sound pressure for different positions and frequencies:

$$\hat{\mathbf{P}} = \begin{pmatrix} \hat{p}_{0,0} & \cdots & \hat{p}_{0,N_m-1} \\ \vdots & \ddots & \vdots \\ \hat{p}_{N_f-1,0} & \cdots & \hat{p}_{N_f-1,N_m-1} \end{pmatrix} \in \mathbb{C}^{N_f\times N_m}, \tag{6}$$

where the notation $\hat{p}_{f,m} = \hat{p}_f(\mathbf{x})|_{\mathbf{x}=\mathbf{x}_m}$ was used. In particular, each column of $\hat{\mathbf{P}}$ can be thought of as a N_f-point Discrete Fourier Transform (DFT) of a discrete-time signal. Similarly, the matrix $\hat{\mathbf{W}} \in \mathbb{C}^{N_f\times N_w}$ contains the weights $\hat{w}_{f,l}$, where the l-th column has the weights of the l-th plane wave for the different frequencies. Notice that the weights and the sound pressures can be transformed into temporal signals using an inverse DFT. The linear operator $D_{pw} : \mathbb{C}^{N_f\times N_w} \to \mathbb{C}^{N_f\times N_m}$ maps these weights $\hat{w}_{f,l}$ to the complex sound pressures $\hat{p}_{f,m}$ using (4) and actually represents a dictionary of N_w plane waves with N_f frequencies. This linear operator is separable for each row of $\hat{\mathbf{P}}$ and $\hat{\mathbf{W}}$ since every frequency is independent. Moreover, each column of these matrices is Hermitian symmetric, and this redundancy can be exploited during evaluation to reduce the computational cost. When the weights $\hat{w}_{f,l}$ are given, (5) represents the *forward problem*. Nevertheless, typically the weights are unknown and an *inverse problem* must be solved to estimate them.

III. Time-domain equivalent source method

Following the same logic of the previous section, a similar time-domain approximation of the sound field is proposed. The time-domain Green's function is defined as

$$\phi_l(\mathbf{x},t) = \frac{1}{4\pi d_l}\delta\left(t - \frac{d_l}{c}\right), \tag{7}$$

where δ is the Dirac delta function and $d_l = \|\mathbf{x}_l - \mathbf{x}\|_2$ is the distance between $\mathbf{x}$ and $\mathbf{x}_l$. In the following $\phi_{l,m}(t)$ will indicate $\phi_l(\mathbf{x},t)|_{\mathbf{x}=\mathbf{x}_m}$. The time-domain Green's function is a particular solution of the non-homogeneous wave equation [34]

$$\nabla^2 p(\mathbf{x},t) - \frac{1}{c^2}\frac{\partial^2 p(\mathbf{x},t)}{\partial t^2} = \delta(\mathbf{x}-\mathbf{x}_l,t) \text{ on } \mathbb{R}^{3\times 1}, \tag{8}$$

with null initial conditions and unbounded domain. The wave equation is the time domain counterpart of Helmholtz equation (3). The sound pressure $p(\mathbf{x},t)|_{\mathbf{x}=\mathbf{x}_m}$ evaluated at a given point in space $\mathbf{x}_m$ is equivalent to a RIR between $\mathbf{x}_l$ and $\mathbf{x}_m$, and $\delta(\mathbf{x}-\mathbf{x}_l,t)$ describes a point source positioned at $\mathbf{x}_l$ that emits a pulse at time $t = 0$. Equation (7) represents what is referred here as an *equivalent source*, i.e. a sound source that emits a spherical wave. In an unbounded or anechoic domain, an equivalent source positioned at $\mathbf{x}_l$ generates a spherical wave which arrives at a position $\mathbf{x}_m$ after traveling the distance $d_{l,m} = \|\mathbf{x}_l - \mathbf{x}_m\|_2$. Similarly to Section II, where the index l indicates the l-th plane wave $\hat{\phi}_{f,l,m}$ with direction $\mathbf{n}_l$, here this index refers to the l-th spherical wave $\phi_{l,m}$ generated at $\mathbf{x}_l$ also arriving to the microphone at a position $\mathbf{x}_m$ from an l-th specific direction. In addition, here a microphone positioned at $\mathbf{x}_m$ would capture a pulse delayed by $d_{l,m}/c$ seconds and attenuated by a factor of $1/(4\pi d_{l,m})$ due to the spherical spreading of energy. The equivalent source can be discretized over time at a sampling frequency F_s using a *fractional delay* filter with Impulse Response (IR) $h_{l,m}$:

$$\phi_{l,m}(n) = \frac{1}{4\pi d_{l,m}}h_{l,m}(n), \tag{9}$$

where now n is the discrete time index. Here, the fractional delay filters are evaluated using Thiran all pass filters, which consist of Infinite Impulse Response (IIR) filters [35]. Spherical waves can be thought of as a generalization of plane waves [12], i.e. when $d_{l,m}$ is large, spherical waves can effectively represent plane waves. Therefore also a finite sum of equivalent sources in the far field can approximate well any sound field in a source-free volume Ω, under the same assumptions as in the previous section. As a consequence, the acoustic model presented in this section can be seen as a generalization of the PWDM. The following expression shows this time-domain acoustic model which is referred to as the Time-domain Equivalent Source Method (TESM):

$$p(\mathbf{x}, n)|_{\mathbf{x}=\mathbf{x}_m} \approx \sum_{l=0}^{N_w-1} \delta(n) * \phi_{l,m}(n) * w_l(n), \tag{10}$$

for $\mathbf{x}_m \in \Omega$. Here $w_l(n)$ is signal of dimension N_t that *weights* the equivalent sources through linear convolution (represented here with the symbol $*$). This signal will be referred as the *weight signal*. If the anechoic source signal that generates the sound field is known, this signal, $s(n)$, can be included in the TESM to increase the quality of the RIR interpolation:

$$p(\mathbf{x}, n)|_{\mathbf{x}=\mathbf{x}_m} \approx \sum_{l=0}^{N_w-1} s(n) * \phi_{l,m}(n) * w_l(n), \tag{11}$$

for $\mathbf{x}_m \in \Omega$. The signal $s(n)$ could correspond to the IR of a loudspeaker measured in an anechoic chamber. When this approximation is used it will be referred to as Source-aware Time-domain Equivalent Source Method (sTESM). The abbreviation (s)TESM will be used to indicate TESM and sTESM simultaneously.

Similar to (5) in Section II, (10) and (11) can be generalized for N_m discrete positions $\mathbf{x}_m \in \Omega$:

$$\mathbf{P} = D_{(s)t}(\mathbf{W}) \tag{12}$$

where $\mathbf{P} \in \mathbb{R}^{N_t \times N_m}$ is a matrix in which the m-th column is the sound pressure signal $p(\mathbf{x}, n)|_{\mathbf{x}=\mathbf{x}_m}$ and $\mathbf{W} \in \mathbb{R}^{N_t \times N_w}$ is a matrix in which the l-th column is the weight signal $w_l(n)$. Therefore the linear operator $D_{(s)t} : \mathbb{R}^{N_t \times N_w} \to \mathbb{R}^{N_t \times N_m}$ maps the weight signals to the sound pressures and represents a dictionary of equivalent sources. This can be computed simply by using (9) with (10) or (11) for $l = 0, \ldots, N_w - 1$ and for $m = 0, \ldots, N_m - 1$.

Another common approach to evaluate the linear operator $D_{(s)t}$ and also the D_{pw} presented in Section II is to create a linear system of equations by vectorizing $\mathbf{P}$ and $\mathbf{W}$:

$$\mathbf{p} = \mathbf{D}\mathbf{w}, \tag{13}$$

where $\mathbf{p} = \text{vec}(\mathbf{P})$ and $\mathbf{w} = \text{vec}(\mathbf{W})$ with vec() indicating the column-major vectorization operator. This has the advantage that if the linear operator has to be evaluated several times, as when solving optimization problems, the fractional delays appearing in (9) or the plane waves appearing in (2) can be computed only once and stored in the matrix $\mathbf{D}$. Nevertheless, looking at the dimensions of $\mathbf{D}$, $N_t N_m \times N_t N_w$, it is clear that storing such a matrix can easily become intractable. As an example, for TESM, consider a time window of 0.1 seconds sampled with $F_s = 8$ kHz ($N_t = 800$). If $N_w = 400$ and $N_m = 12$, $\mathbf{D}$ is a $96 \cdot 10^2 \times 32 \cdot 10^4$ matrix which for a 64 floating point format would result in a memory requirement of approximately 24.6 GB. In such cases, it will be necessary to sacrifice computational power for memory by directly evaluating the linear operators when needed. This strategy will lead to *matrix-free* optimization.

IV. The inverse problem

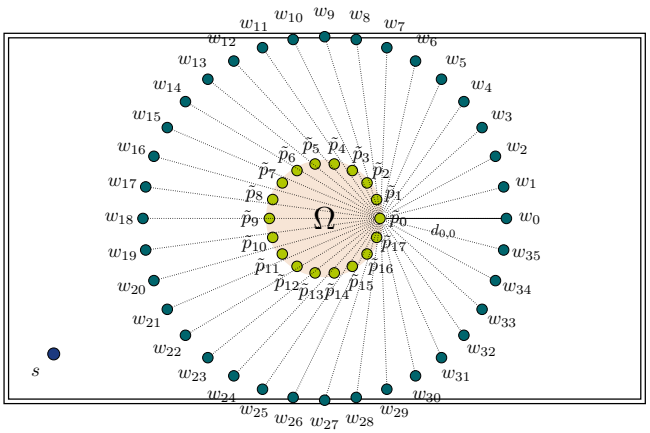

Figure 1. A cross-section of a room viewed from above. A sound source placed near the front left corner creates a reverberant sound field that is captured by a spherical microphone array (light green dots). The sound pressure is then matched using a set of equivalent sources $\phi_{l,m}$ (dark green dots). This makes it possible to perform RIR interpolation over the shaded volume Ω surrounded by the microphones. The distance $d_{0,m}$ between the first microphone and the equivalent sources is shown with dotted lines.

Figure 1 shows the measurement set-up needed for the RIR interpolation. A far field sound source, represented in the figure in the front left corner of the room, generates a reverberant sound field. A spherical microphone array of N_m microphones (light green dots) is placed in the middle of the room and used to measure the RIRs at these positions. Notice that the framework described this far does not assume any particular geometry for the position of the microphone measurements: it is only assumed that the relative distances between the microphone measurements are known. The choice of a spherical microphone array is arbitrary. The aim is to interpolate these RIRs inside the volume Ω. In this volume, it is assumed that the models presented in Sections II and III can well approximate the sound field. What is sought by the inverse problem is to extrapolate out of the microphone measurements the optimal weight signals that lead to the best sound field approximation possible. This inverse problem can be formulated as an optimization problem:

$$\mathbf{W}^\star = \underset{\mathbf{W}}{\text{argmin}} \quad f(\mathbf{W}) = \frac{1}{2}\|D(\mathbf{W}) - \tilde{\mathbf{P}}\|_F^2, \tag{14}$$

where $\|\cdot\|_F$ is the Frobenius norm defined as

$$\|\mathbf{A}\|_F = \|\text{vec}(\mathbf{A})\|_2. \tag{15}$$

Notice that here the linear operator $D(\cdot)$ has no subscript, which implicitly means that any of the acoustic models pre-

sented in the previous sections can be used. The columns of the matrix $\tilde{\mathbf{P}}$ contain the microphone measurements, i.e. the N_t-long measured RIR signals. The aim of this optimization problem is to minimize the *un-regularized cost function* $f(\mathbf{W})$, which consists of the distance between the measured RIRs and the sound pressure of the acoustic model. However, problem (14) is heavily *ill-posed*: if many spherical (plane) waves are used to construct $D(\cdot)$, multiple solutions for $\mathbf{W}$ can minimize the cost function effectively. This will in general lead to *over-fitting*: the measured RIRs will coincide with the sound pressure of the acoustic model but only at the microphone positions, leading to a poor RIR interpolation. To avoid this, it is necessary to regularize problem (14). A common regularization method is *Tikhonov regularization* which consists of modifying the cost function by adding a regularization term:

$$\mathbf{W}_T^\star = \underset{\mathbf{W}}{\operatorname{argmin}} \quad f(\mathbf{W}) + \frac{\lambda}{2}\|\mathbf{W}\|_F^2. \tag{16}$$

Here the parameter λ controls the level of the regularization and in practice balances the *prior knowledge* induced by the regularization with the information provided by the microphone measurements that appears in $f(\cdot)$. For the case of Tikhonov regularization, this prior knowledge represents the expectation that the components of $\mathbf{W}$ are small. As simulations will show later, this regularization does not produce good results as it does not include any spatial information.

If it is expected that the matrix $\mathbf{W}$ is *sparse* instead, i.e. that the majority of the components of $\mathbf{W}$ are zero, a sparsity promoting regularization can be used, e.g. the l_1*-norm regularization*:

$$\mathbf{W}_1^\star = \underset{\mathbf{W}}{\operatorname{argmin}} \quad f(\mathbf{W}) + \lambda\|\text{vec}(\mathbf{W})\|_1. \tag{17}$$

Problem (17) is also known as the Least Absolute Shrinkage and Selection Operator (LASSO) and has been widely used in CS. As opposed to Tikhonov regularization, the l_1-norm regularization, when it is used with the (s)TESM, promotes *spatio-temporal sparsity*. This becomes clear by looking at the structure of $\mathbf{W}$, which in the case of (s)TESM contains the time-domain weight signals that control equivalent sources placed at different positions. Spatio-temporal sparsity can be justified physically: in a reverberant sound field generated by an impulsive source, sound waves arrive from specific directions at specific times. On the other hand, when the l_1-norm regularization is used with the PWDM, *spatio-spectral sparsity* is promoted. Once more this becomes clear by looking at the structure of $\hat{\mathbf{W}}$, although a physical interpretation becomes more difficult. As simulations will show, the l_1-norm regularization does indeed produce good results for the (s)TESM and moderate results for the PWDM.

Another possible regularization is the *sum of* l_2*-norms regularization*:

$$\mathbf{W}_{\Sigma l_2}^\star = \underset{\mathbf{W}}{\operatorname{argmin}} \quad f(\mathbf{W}) + \lambda \sum_{l=0}^{N_w-1} \|\mathbf{W}_{:,l}\|_2. \tag{18}$$

where $\mathbf{W}_{:,r}$ indicates the rth column of $\mathbf{W}$. This regularization is aimed to have only few columns of $\mathbf{W}$ to have non-zero coefficients. In the PWDM and (s)TESM case this regularization imposes solely *spatial sparsity*, and it is a special case of *group sparsity*. When the l_1-norm regularization is used with the PWDM, the optimization problem becomes fully separable. Since all frequencies are independent in the linear operator of the PWDM and the same is true for the l_1-norm regularization it is possible to cast individual optimization problems per frequency. This property is no longer valid for sum of l_2-norms regularization, which implies that the optimization problem is not separable anymore. Despite this disadvantage, simulations will show that the sum of l_2-norms regularization outperforms the l_1-norm one when applied to PWDM. Clearly, when a wide band sound source generates the sound field, imposing sparsity in the frequency domain is not a good choice and therefore spatial sparsity provides better results.

Finally, an important aspect is the choice of the equivalent source positions in (s)TESM. In a reverberant environment, sound waves can arrive from every direction but the dictionaries must be of finite dimension. As Fig. 1 shows, for the 2-dimensional case the equivalent sources can be placed uniformly on a circle (dark green dots). Nevertheless, in 3-dimensions the uniform sampling of the surface of a sphere is not unique and for instance sampling uniformly the azimuthal and the polar angles leads to a concentration of equivalent sources near the poles. In order to avoid this, Fibonacci lattices, which provide nearly uniform sampling of the surface of a sphere, are used [36], [37]. The azimuthal and polar angles obtained with this lattice can be used also for the candidate directions of the plane waves in the PWDM.

V. Optimization algorithm

In this section firstly the FBS algorithm is described. Secondly, the quasi-Newton accelerated FBS is briefly presented. Further details about these methods can be found in [**?**], [29], [31].

The FBS generalizes the well-known gradient descent algorithm to a class of non-smooth cost functions. The FBS can be used to solve optimization problems with the following structure:

$$\mathbf{W}^\star = \underset{\mathbf{W}}{\operatorname{argmin}} f(\mathbf{W}) + g(\mathbf{W}), \tag{19}$$

where $f(\cdot)$ is convex and smooth, such as the un-regularized cost function (14), and $g(\cdot)$ can be non-smooth, such as the l_1-norm in (17). Clearly, all of the optimization problems presented in Section IV have cost functions which can be split in such a fashion.

Starting from an initial guess $\mathbf{W}^0$, one can iterate the expression

$$\mathbf{W}^{k+1} = T_\gamma(\mathbf{W}^k) = \text{prox}_{\gamma g}\left(\mathbf{W}^k - \gamma\nabla f(\mathbf{W}^k)\right), \tag{20}$$

to obtain a solution. Here $\nabla f(\cdot)$ is the Jacobian operator of $f(\cdot)$, $T_\gamma(\cdot)$ is the *forward-backward operator* and $\text{prox}_{\gamma g}(\cdot)$ is the *proximal mapping* of the function $g(\cdot)$, defined as:

$$\text{prox}_{\gamma g}(\mathbf{W}) = \underset{\mathbf{U}}{\operatorname{argmin}} \frac{1}{2\gamma}\|\mathbf{U}-\mathbf{W}\|_F^2 + g(\mathbf{U}). \tag{21}$$

	l_1-norm	$\sum$ l_2-norms	Tikhonov
$g(\mathbf{W})$	$\lambda\|\text{vec}(\mathbf{W})\|_1$	$\lambda\sum_{l=0}^{N_w-1}\|\mathbf{W}_{:,l}\|_2$	$\frac{\lambda}{2}\|\mathbf{W}\|_F^2$
$\text{prox}_{\gamma g}(\mathbf{W})$	$\text{sign}(\mathbf{W})\max(\mathbf{0},\|\mathbf{W}\|-\gamma\lambda)$	$\max(0,1-\gamma\lambda/\|\mathbf{W}_{:,l}\|_2)\mathbf{W}_{:,l}$ for $l=0\dots N_w-1$	$\frac{1}{\gamma\lambda+1}\mathbf{W}$
λ_{max}	$\|\text{vec}(D^a(\mathbf{P}))\|_\infty$	$\left\|\left[\|D^a(\mathbf{P})_{:,0}\|_2,\dots,\|D^a(\mathbf{P})_{:,N_w-1}\|_2\right]^T\right\|_\infty$	$\|D^a(\mathbf{P})\|_2^2$

Table I
TABLE SHOWING THE DIFFERENT REGULARIZATIONS USED IN THIS PAPER AND THEIR EQUIVALENT PROXIMAL MAPPINGS. THE LAST ROW SHOWS THE MAXIMUM VALUE OF λ USED FOR THE DIFFERENT REGULARIZATIONS.

The argument inside the parenthesis of $\text{prox}_{\gamma g}(\cdot)$ in (20) corresponds to the gradient descent iteration for the un-regularized cost function $f(\cdot)$. Therefore a simple interpretation of (20) is the following: a new iterate is obtained using the gradient descent for $f(\cdot)$ (forward step)

$$\mathbf{W}_{gd}^{k+1} = \mathbf{W}^k - \gamma\nabla f(\mathbf{W}^k), \tag{22}$$

and then a second optimization problem, cf. (21), is solved to project the iterate $\mathbf{W}_{gd}^{k+1}$ into a modified solution *close* to the minimum of $g(\cdot)$ (backward step). The scalar γ is the *step-size* which must be sufficiently small to ensure convergence. One of the main advantages of the FBS algorithm is that often the proximal mapping can be computed very cheaply. For example, when $g(\cdot)=\lambda\|\cdot\|_1$ the proximal mapping (21) has an analytical solution and reduces to a soft-thresholding of the elements of $\mathbf{W}_{gd}^{k+1}$ [29]. Table I summarizes the proximal mappings for the different regularizations used in this paper which all have analytical solutions and are cheap to compute. As stated earlier, a solution to (19) can be obtained by iterating (20) leading to a simple algorithm with a convergence speed comparable to the gradient descent algorithm.

In [31] a novel algorithm that dramatically accelerates the FBS algorithm has been proposed. The idea is to modify the FBS with the following iterations:

$$\mathbf{W}^{k+1} = T_\gamma(\mathbf{W}^k) + \tau\mathbf{S}^k, \tag{23}$$

where $\mathbf{S}^k$ is a corrective direction and τ is a step-size. It can be proven that the optimality condition for $\mathbf{W}^\star$ to be the optimal solution of (19) is given as [31]

$$R_\gamma(\mathbf{W}^\star) = \mathbf{W}^\star - T_\gamma(\mathbf{W}^\star) = \mathbf{0}, \tag{24}$$

where $R_\gamma(\cdot)$ represents the *fixed-point residual*. At each iteration, the corrective direction $\mathbf{S}^k$ is computed using curvature information of the fixed-point residual obtained through quasi-Newton Limited memory BFGS (L-BFGS) updates to accelerate convergence. The step-sizes γ and τ are chosen adaptively with two separate line-search procedures to ensure global convergence. These line-search procedures only rely on the very same predictions of the FBS which, together with the L-BFGS updates, have minimal memory requirement. Further details on the algorithm are left out here for brevity and the interested reader can find more information in [31]. An implementation of the algorithm is also available online [38], [39].

VI. COMPUTATION OF THE JACOBIAN

The Jacobian $\nabla f(\cdot)$ appears in the FBS iterations (20), and consequently also in the accelerated version (23). Moreover the un-regularized cost function $f(\cdot)$ must also be computed as it is needed in the line-search procedures. The efficient computation of $f(\cdot)$ and $\nabla f(\cdot)$ is therefore crucial since these are required in every iteration of the optimization algorithm.

Regarding the computation of $f(\cdot)$, by inspecting its definition

$$f(\mathbf{W}) = \frac{1}{2}\|\underbrace{D(\mathbf{W})-\tilde{\mathbf{P}}}_{\mathbf{R}}\|_F^2, \tag{25}$$

it can be noticed that the most computationally expensive operation lies in the evaluation of $D(\mathbf{W})$. Here the residual $\mathbf{R}$ is the difference between the measured sound pressure and the sound pressure provided by the acoustic model of choice. It was already mentioned at the end of Section III how avoiding the storage of $D(\cdot)$ into a matrix $\mathbf{D}$ and using a recursive computation instead can be beneficial for a large scale optimization problem, as it minimizes the memory requirements.

The same strategy will be used in the computation of the Jacobian $\nabla f(\cdot)$. The Jacobian can be written as:

$$\nabla f(\mathbf{W}) = D^a(D(\mathbf{W})-\tilde{\mathbf{P}}) \Rightarrow \mathbf{J} = D^a(\mathbf{R}), \tag{26}$$

where $D^a(\cdot):\mathbb{U}^{N_mN_t\times N_wN_t}\rightarrow\mathbb{U}^{N_wN_t\times N_mN_t}$ is the *adjoint operator* [33] of $D(\cdot)$, $\mathbf{J}:\mathbb{U}^{N_wN_t\times N_mN_t}$ is the Jacobian matrix and $\mathbb{U}$ is either $\mathbb{R}$ or $\mathbb{C}$. If the linear operator $D(\cdot)$ consisted of a matrix multiplication the adjoint operator would have been the conjugate-transpose operator of that matrix. It can be noticed that the adjoint operator is applied to the residual $\mathbf{R}$ which is readily available after the computation of $f(\cdot)$.

For the (s)TESM the linear operator $D_{(s)t}(\cdot)$ is computed by recursively convolving the weight signals $w_l(n)$ with $(s(n))\ *\ \phi_{l,m}(n)$ as shown in (10) and (11). Since the adjoint operator of convolution is the cross correlation [40], the lth column of the Jacobian matrix $\mathbf{J}$ can be computed as:

$$j_l(n) = \sum_{m=0}^{N_m-1}(s(n)*\phi_{l,m}(n))\odot r_m(n) \tag{27}$$

where $\odot$ indicates the cross correlation operator and $r_m(n)$ is the residual signal of the m-th microphone measurement, i.e. the m-th column of $\mathbf{R}$. This operation can be repeated iteratively to compute $\mathbf{J}$ as in the case of the evaluation of $D_{(s)t}(\cdot)$.

Analyzing (27) from a different perspective, the adjoint operator can be thought of as a new swapped forward problem

with N_m equivalent sources positioned at $\mathbf{x}_m$ having as weight signals the time reversed residual $r_m(-n)$. This sound field is then captured by N_w microphone measurements positioned at $\mathbf{x}_l$. Equation (27) is in fact equivalent to:

$$j_l(-n) = \sum_{m=0}^{N_m-1} (s(n) * \phi_{l,m}(n)) * r_m(-n) \qquad (28)$$

which in practice is used in the computation since it requires the very same fractional delay filters as the ones used in $D_{(s)t}(\cdot)$.

Similarly, for the PWDM the adjoint operator can be obtained using the following equation:

$$\hat{j}_{f,m} = \sum_{m=0}^{N_m-1} \hat{r}_{f,l} \hat{\phi}^*_{f,l,m}, \qquad (29)$$

where $\hat{j}_{f,m}$ is the (f,m)-th element of the complex Jacobian matrix $\hat{\mathbf{J}}$, $\hat{r}_{f,l}$ is the (f,l)-th element of the residual $\hat{\mathbf{R}}$ and with * denoting the complex conjugate operation.

VII. Simulation results

In this section the RIR interpolation performance is analyzed using the different acoustic models described in Sections II and III and the different regularizations presented in Section IV. The microphone signals used in the inverse problem are generated using the Randomized Image Method (RIM), a modified version of the IM that avoids the presence of sweeping-echos in the simulated RIRs [41]. As Fig.1 shows, the reverberant acoustic environment consists of a box-shaped room with dimensions $[L_x, L_y, L_z] = [6, 3.5, 4]$ m. A sampling frequency of $F_s = 8$ kHz is used with a time window of 70 ms ($N_t = N_f = 560$).

Frequency dependent impedances are used for the walls: this is achieved by using IIR filters for the image sources which are then self-convolved for the higher order image sources. These filters are obtained from measured absorption coefficients found in [42] using the same procedure described in [43], i.e. optimizing the coefficients of the IIR filters via a damped Gauss-Newton method. The fractional delays of the RIM are modeled using Finite Impulse Response (FIR) filters as in [44]. Software is available at [45]. Only one material is used to model all of the room acoustic impedances. The resulting sound field has a spatially averaged reverberation time of $T_{30} = 0.065$ s.

An omnidirectional sound source is placed at $\mathbf{x}_s = [0.75, 0.43, 2]$ m producing a source signal $s(n)$, an impulse filtered using a bandpass 4th order Butterworth filter from 20 Hz to 3.2 kHz. Microphones are placed on a spherical array of radius 0.35 m with center in the middle of the room at $\mathbf{x}_c = [L_x, L_y, L_z]/2$. The microphone positions form a Fibonacci lattice. All the microphone signals are corrupted with additive white noise with a SNR = 15 dB. When comparing performances with different regularizations and acoustic models the same noise is added. Such a high SNR is motivated by the fact that it is assumed that most of the noise is reduced in post-processing by averaging the RIR microphone measurements. The RIRs are interpolated in the volume (0.18 m^3) enclosed by the spherical microphone array. Nevertheless, since it is not possible to compute the true RIR at every position inside the volume, their quality is evaluated on an *interpolation volume*, that is a cuboid volume of dimensions $[0.43, 0.43, 0.63]$ m which is spatially uniformly sampled ($N_{in} = 300$). and placed inside the spherical microphone array.

For the (s)TESM $N_w = 700$ equivalent sources are positioned in a Fibonacci lattice of radius 1.75 m centered at $\mathbf{x}_c$. Similarly, for the PWDM, $N_w = 700$ directions are used and these are equivalent to those used in the (s)TESM. The number of plane wave directions and equivalent sources was obtained empirically. Simulations were performed also for different configurations of source position, room dimensions and reverberation time, which led to similar results as the ones presented in the following subsections and are therefore omitted here for brevity. Notice that all of the simulations presented here are reproducible [46].

A. Choice of regularization parameter λ

As described in Section IV, the parameter λ controls the level of the regularization. The regularization can be viewed as imposing a particular prior knowledge on $\mathbf{W}$. This prior knowledge is not always available and often one wants to extrapolate it out of the available microphone measurements. This requires finding the best λ which controls the balance between how much a model fits the available microphone measurements and how important the prior knowledge is. The optimal balance will simultaneously avoid over-fitting ($\lambda \to 0$) and predictions based on pure inference (large λ), which both produce poor results.

In this paper, λ is tuned using K-fold Cross Validation (KCV) [47] in which scheme the available measurements are split into K folds, namely K different groups of equal size. Here these folds consist of K scrambled groups of measured RIRs. For a given λ, the acoustic model is *trained* using only $K-1$ folds by solving the inverse problem using a particular regularization. The fold left outside is used to test the training performance and to produce the cross validation error:

$$\epsilon_{cv} = \frac{1}{N_{cv}} \sum_{m=1}^{N_{cv}} \|\mathbf{P}_{:,m} - \tilde{\mathbf{P}}_{:,m}\|_2^2 / \|\mathbf{P}_{:,m}\|_2^2, \qquad (30)$$

i.e. the average of the Normalized Mean Squared Error (NMSE) of the N_{cv} RIRs belonging to the fold not used in the training. Here $\mathbf{P}$ is reconstructed from the optimal solution $\mathbf{W}^\star$ by means of either (5) or (12). This procedure is repeated K times, each time leaving out a different fold to produce a new cross validation error. The K cross validation errors are further averaged to produce $\bar{\epsilon}_{cv}$, the averaged cross validation error. A set of values for λ is tested using this procedure and the λ that gives the minimum averaged cross validation error is then chosen. Finally, the model is trained using all of the available microphone measurements with the chosen λ.

The KCV starts with an over-regularized inverse problem, with $\lambda = \lambda_{\text{max}}$, and continues by logarithmically decreasing λ.

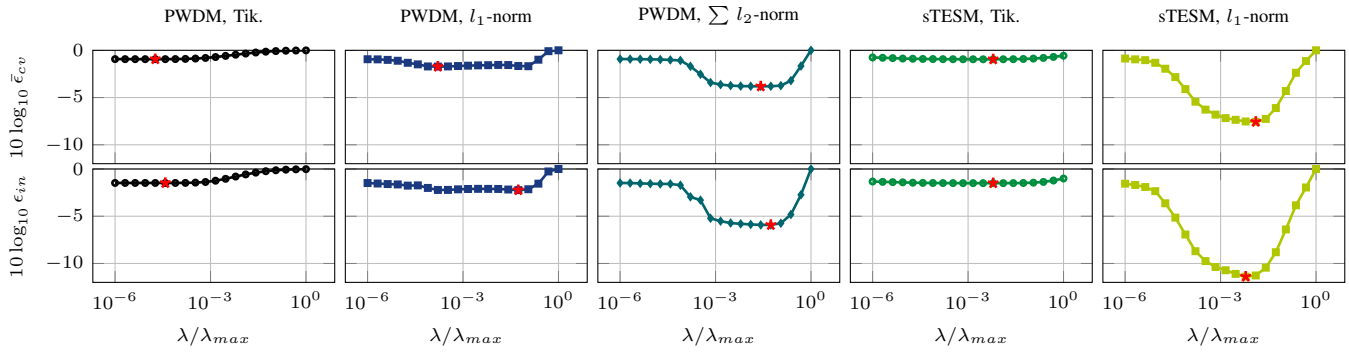

Figure 2. Simulated averaged cross validation error curve (top plots) and interpolation error curve (bottom tops) as a function of λ (normalized by $\lambda_{\max}$) for different types of acoustic models and regularizations using $N_m = 12$ microphones.

Here $\lambda_{\max}$ is chosen as the maximum value for which the solution of the inverse problem is null. Table I, reports the values of $\lambda_{\max}$ for each regularization [29], [48]. Notice that, for Tikhonov regularization, the solution is not null for $\lambda < \infty$, however the $\lambda_{\max}$ given in the table gives still an over-regularized solution.

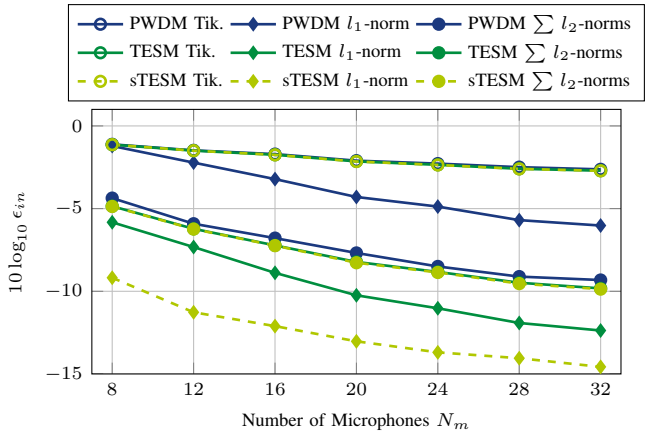

Figure 3. Simulated interpolation error for different types of acoustic models and regularizations as a function of the number of microphones.

In a simulated environment, it is actually possible to check if the KCV indeed returns the best λ since the *interpolation error* ϵ_{in} can be computed, i.e. the average of the NMSE between the true RIRs and the interpolated RIRs, which in many real scenarios would not be available. Specifically, the interpolation error is computed using the same formula (30) on the interpolation volume. Figure 2 shows the averaged cross validation error and the interpolation error for $N_m = 12$ microphones as a function of λ for different types of regularization and acoustic models. Here and in the following a $K = 4$ KCV is used. In general the minima of the errors coincide and when they do not they are either very close to each other or they belong to flat regions of the error curves. This condition was verified for all the results presented here and in the following, the only exception being when $N_m = 4$ microphones are used. For this reason, these results are not shown. In conclusion, the results show that KCV is a good strategy for tuning λ if sufficient microphone measurements are available.

B. Comparison between acoustic models

In Fig. 3 the interpolation error is shown as a function of the number of microphone measurements used in the spherical array, while in Fig. 4 the worst-case interpolated RIR and its equivalent DFT are shown. Here worst-case indicates the position with maximum error in the interpolation volume. Looking at Fig. 3, it is clear that Tikhonov regularization gives almost equivalent performances with either PWDM or (s)TESM. This is expected, as discussed in Section IV, since Tikhonov regularization simply avoids large energy of the weight signals without imposing any spatial information. Here only the low frequencies are correctly interpolated as Fig. 4(a) shows. An improvement is achieved with the PWDM and l_1-norm regularization where spatio-spectral sparsity is promoted. In Fig. 4, it can be seen that the direct component becomes more pronounced but many artifacts are present. Clearly, as Fig. 3 shows, promoting sparsity in the frequency domain when the sound field is generated by a wide band signal is not a good choice. When only spatial sparsity is promoted, using the sum of l_2-norms regularization, better results are achieved. With this type of regularization all the acoustic models show similar performance: the TESM and the sTESM are practically equivalent while the interpolation errors of the PWDM are on average 0.45 dB higher with respect to the (s)TESM. Finally, all the results discussed so far are outperformed using the (s)TESM and promoting spatio-temporal sparsity with the l_1-norm regularization. In Fig. 3, it can be seen that the same interpolation error obtained with $N_m = 32$ microphones and spatio-spectral sparsity (PWDM with l_1-norm) can be achieved with only $N_m = 8$ microphones using the TESM and spatio-temporal sparsity. A similar comparison can be made between the PWDM with spatial sparsity that reaches -9.3 dB interpolation error with $N_m = 32$ microphones and the sTESM that reaches -9.2 dB using only $N_m = 8$ microphones. When comparing the TESM with respect to the sTESM instead, it can be noticed that the sTESM significantly outperforms the TESM, particularly when the number of microphone measurements is smaller as Fig. 3 shows. Comparing Fig. 4(d) with Fig. 4(e) it can be noticed that the sTESM provides a more accurate representation of the

	PWDM, Tik.	PWDM, l_1-norm	PWDM $\sum$ l_2-norms	TESM, l_1-norm	sTESM, l_1-norm
NMSE (dB)	-0.58	-0.93	-4.42	-6.14	-9.37
NMSE (dB) Early Reflections	-0.52	-1	-4.5	-6.37	-9.8
NMSE (dB) Late Reverberation	-2.57	1	-1.88	-1.71	-3.39

Table II
SIMULATED WORST-CASE NMSES OF THE INTERPOLATED RIRS SHOWN IN FIG. 4. THE NMSES ARE ALSO SHOWN FOR THE EARLY REFLECTIONS OF THE RIRS (UP TO 30 ms) AND THE LATE REVERBERATION (STARTING FROM 30 ms).

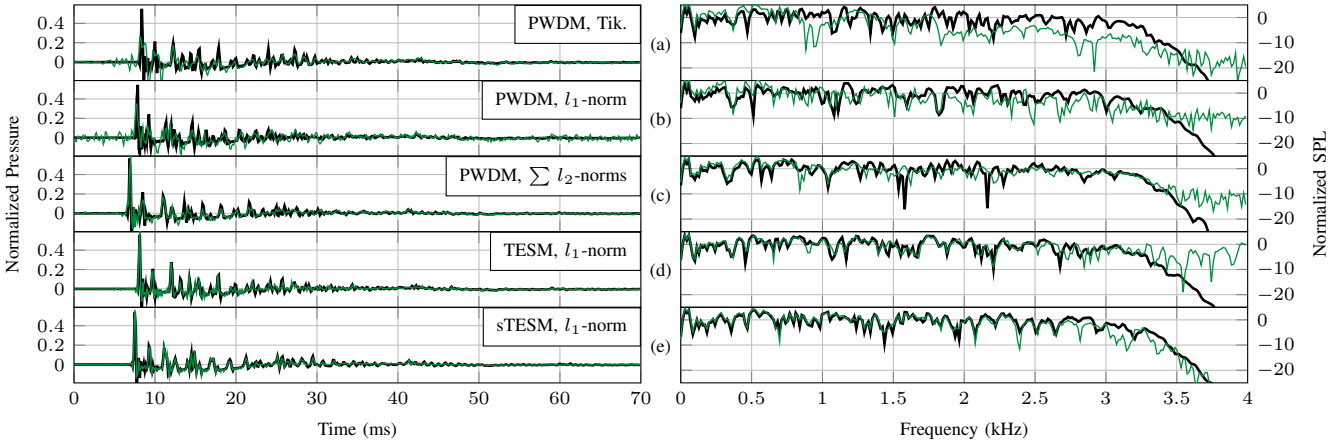

Figure 4. Simulated worst-case RIR interpolation for different types of acoustic models and regularizations using $N_m = 12$ microphones. Black thick line shows the original RIR while the green thin line shows the interpolated RIR.

high frequencies. This is due to the fact that the equivalent sources of the sTESM are already *shaped* with the spectrum of $s(n)$.

Finally, looking at Fig. 4 it is interesting to compare the RIR interpolation accuracy for the early reflections and late reverberation. Table II summarizes the worst-case NMSEs between the interpolated and the original RIRs appearing in Fig. 4. Here, in the last two rows, the NMSEs of the early reflections and late reverberation are computed separately. In general early reflections are interpolated more accurately than late reverberation for all the different types of regularizations and acoustic models except for the case of the PWDM with Tikhonov regularization, where the opposite happens. It can be seen that despite the higher reflection density in the late part of the RIRs, the sTESM with spatio-temporal sparsity still achieves the best results. Nevertheless it is difficult to conclude which regularization may achieve the best accuracy for the late reverberation. The early reflections constitute the most energetic part of the RIRs and therefore the unregularized cost function $f(\cdot)$ is unbalanced towards them. Using a weighted norm in $f(\cdot)$ would balance the fitting and different regularizations could be used for different parts of the RIRs as well, with the disadvantage of complicating the regularization tuning procedure. A thorough analysis of these matters goes beyond the scope of this paper and is left for future work.

C. Analysis of weight signals

The weight signals are not only useful to perform the RIR interpolation but can also be used to provide a novel spatio-temporal visualization of the reverberant sound field, similar to the one proposed in [49]. It is important to outline that compared to the visualizations of [49], here a wider volume is represented instead i.e. the volume where the RIR interpolation is performed. Figure 5 shows such a visualization, where for each regularization and acoustic model the weight signals are plotted simultaneously using spherical coordinates: the radius indicates the absolute value $20\log_{10}|w_l(n)|$ of the l-th weight signal (in dB) with its corresponding specific direction identified by the azimuthal and polar angle. Time is represented with color, with lines becoming darker and thinner as time proceeds. This enables one to view the direction of arrival of a specific reflection at a specific time. Moreover the signals are normalized by the maximum absolute value of $\mathbf{W}^\star$. In all the figures a single line, the thickest and with lighter color, reaches 0 dB in the direction associated to the sound source, indicating that all the methods are capable of reconstructing the line-of-sight component correctly. In the left hand figures, these lines point to $\theta \approx 90°$ indicating that the sound source has the same height as the center of the microphone array. In the right hand figures instead, these lines point to $\varphi \approx 220°$ showing that the sound source is located in the front left corner of the room as shown in Fig. 1. Clearly, Tikhonov regularization manages to predict only the line-of-sight component as Fig. 5(a) shows: reflections keep arriving from all directions as time proceeds. This is not the case for the other results shown here, where the direction of arrival of higher order reflections can be clearly distinguished. In the left hand figures, where the weight signals are plotted as a function of the polar angle θ, it is possible to clearly identify the strongest reflections coming from the ceiling ($\theta \approx 45°$), walls ($\theta \approx 90°$) and floor ($\theta \approx 150°$). On the other hand,

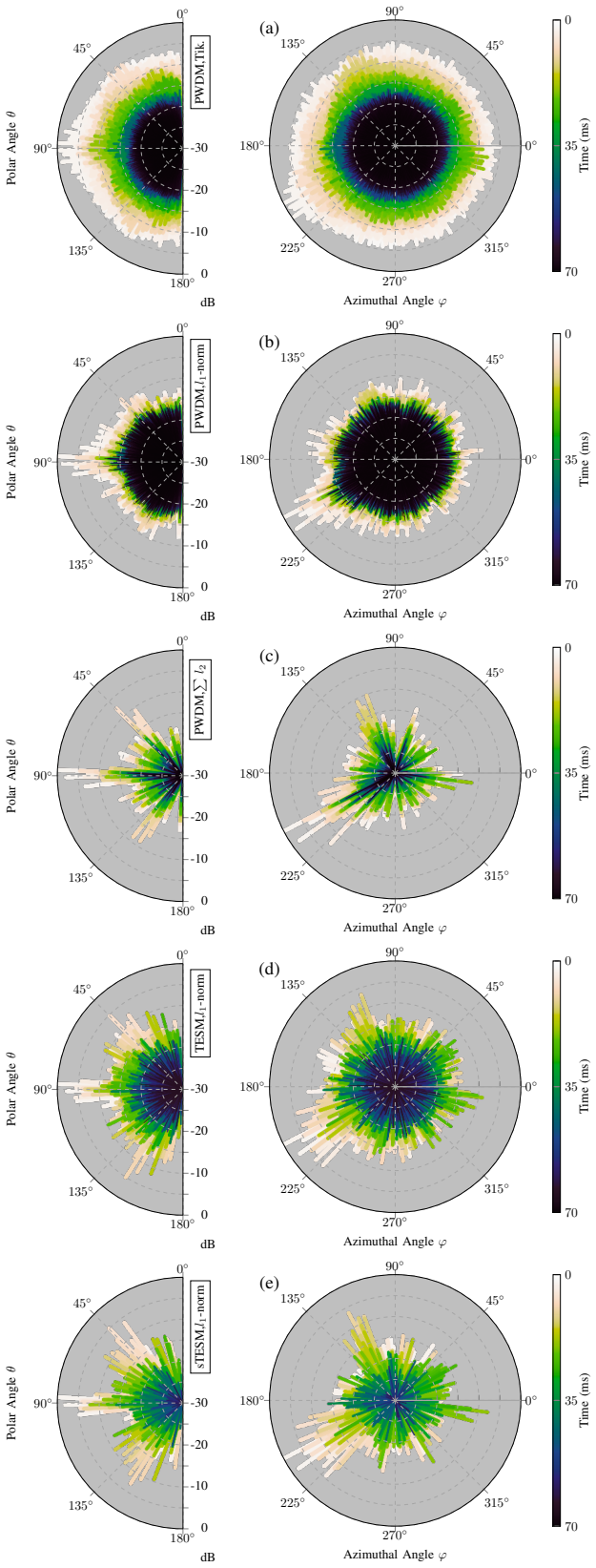

Figure 5. Visualization of the weight signals for different types of acoustic models and regularizations using $N_m = 12$ microphones. The radius in dB indicates the absolute value of the weight signal for a given direction. Directions are shown using spherical coordinates. Color indicates time: as time proceeds darker colors and smaller linewidth are used to plot the weight signals. Animated versions of these figures can be found at the following links: ftp://ftp.esat.kuleuven.be/stadius/nantonel/Videos/RIR_Intp/ and [46].

the right hand figures give a top view of the reflections. By combining each pair of figures and looking at the radius and color of the lines, one can understand the direction and time of arrival of a specific reflection. Notice that there is a similar trend as observed in the analysis of the performance of the RIR interpolations, with spatial and spatio-temporal sparsity giving clearer spatial information. For the PWDM, the weight signals obtained with spatio-spectral sparsity, Fig. 5(b), never decay completely to zero as time proceeds, while this is not the case for Fig. 5(c) where spatial sparsity minimizes the number of directions. Here one can clearly distinguish between different directions of arrival of higher order reflections, particularly the first order reflections coming from the walls, ceiling and floor. The same can be stated for the weight signals of (s)TESM with sTESM having much sparser signals with respect to TESM, as expected. Although many more directions of arrival are present in the (s)TESM solutions when compared to the PWDM solution with spatial sparsity, this does not necessarily mean that the PWDM solution is sparser. In fact, in the latter case, having many directions of arrival goes against the regularization which encourages grouping different reflections with similar directions of arrival and bundles them into a single one. This is obviously not beneficial for either the RIR interpolation or for the spatio-temporal visualization. Finally, the level of sparsity of the weight signals obtained with sTESM is maximal when compared with the other results. In the example shown in Fig. 5, for $N_m = 12$, the sTESM matrix $\mathbf{W}^\star$ has only 0.4% of its components non-zero, while TESM has 1.4%. The PWDM with spatial sparsity has 29.14% both in the frequency and time domain: despite the appearance of Fig. 5(c) the active weight signals decay over time to low levels but are always non-zero. On the other hand, PWDM with spatio-spectral sparsity has a dense matrix in the time-domain, while its frequency domain counterpart has 27% of its components non-zero.

VIII. Experimental results

In this section experimental results are presented. The RIR interpolation is performed using measured RIRs from the single- and multichannel audio recordings database (SMARD) [50]. The RIRs were measured in a box-shaped listening room with dimensions $[L_x, L_y, L_z] = [7.3, 8.1, 2.9]$ m and a reverberation time of $T_{30} = 0.097$ s. The database provides RIRs measured for different configurations of loudspeakers and microphone arrays. In this experimental study, configuration 1002 was used: here a Brüel & Kjær OmniSource 4295 loudspeaker unit positioned at $\mathbf{x}_s = [2.0,\ 6.5,\ 1.4]$ m and an orthogonal microphone array were used. The orthogonal microphone array consists of 3 uniform linear microphone arrays which share their origins at $\mathbf{x}_m = [4.4,\ 3.1,\ 1.5]$ m and are placed orthogonally to each other. Each uniform linear microphone array has 7 microphones spaced 5 cm apart. Here, only the $N_m = 15$ microphones belonging to the endings of the orthogonal microphone array's branches are used for the training, while the remaining $N_{in} = 6$ microphones closer to the origin are used to compute the interpolation error ϵ_{in} and to tune λ, thus avoiding the need of using a KCV scheme. The

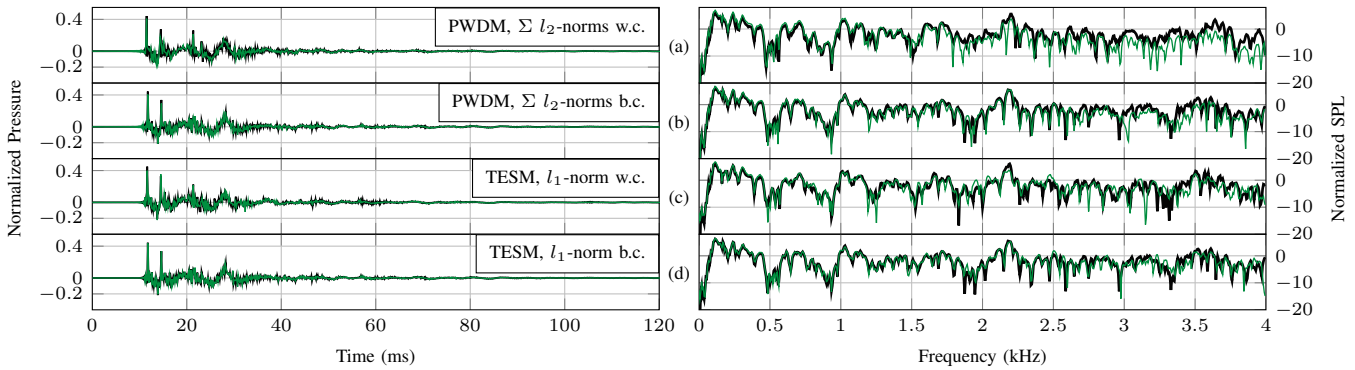

Figure 6. Experimental (w.c. and b.c.) results of RIR interpolation for different types of acoustic models and regularizations using $N_m = 15$ measured RIRs. Black thick line shows the original RIR while the green thin line shows the interpolated RIR. NMSE: (a) -5.5 dB (b) -9.43 dB (c) -7.11 dB (d) -11 dB.

	PWDM, l_1-norm	PWDM $\sum$ l_2-norms	TESM, l_1-norm
ϵ_{in} (dB)	-5.47	-7.22	-8.51

Table III
EXPERIMENTAL INTERPOLATION ERROR FOR DIFFERENT TYPES OF ACOUSTIC MODELS AND REGULARIZATIONS USING 15 MEASURED RIRS.

$N_w = 700$ equivalent sources are positioned in a Fibonacci lattice of radius 2.87 m centered at $\mathbf{x}_m$. All of the RIRs are down-sampled to a sampling frequency of $F_s = 8$ kHz. Note that all of the experiments presented here are reproducible using the code found in [46] and the RIR database available online at [50].

Table III shows the interpolation errors obtained using different types of acoustic models and regularizations. Note that the sTESM cannot be applied since the IR of the loudspeaker is not available. As in the simulation results, the TESM combined with spatio-temporal sparsity achieves the best performance. Looking at Fig. 3, it can be observed that these values are quite close to the ones obtained in the simulation results using $N_m = 16$ microphones. Notably, for the two cases where the PWDM was used, the experimental results are slightly better than the simulation results despite the fact that these were obtained using one microphone less. This is probably due to higher SNR of the experimental RIRs compared to the ones used in the simulations. Finally, Fig. 6 compares the interpolated RIRs with the original ones for the worst-case and best-case results: similarly to Fig. 4, where only the worst-case results are shown, early reflections and low frequencies are reconstructed with higher accuracy with respect to the late reverberation and high frequencies respectively.

IX. CONCLUSIONS

In this paper the problem of spatially interpolating measured RIRs is posed as an inverse problem. Two different acoustic models that are able to approximate any sound field in a source-free volume generated by a far field source are presented and compared: the PWDM which is a widely known frequency domain model and (s)TESM, a novel time-domain method that can also incorporate the knowledge of the source signal generating the sound field. It is shown that various sparsity promoting regularizations can be used to cope with the ill-posed nature of the inverse problem. Spatio-spectrally and spatially sparse solutions can be obtained with the PWDM and with the l_1-norm and sum of l_2-norms regularization respectively, while a spatio-temporally sparse solution can be encouraged using (s)TESM with l_1-norm regularization.

These inverse problems turn out to be large scale optimization problems and so great care must be taken in finding a computationally tractable algorithm. For this reason, the acoustic models are described as linear operators that can be computed with minimal memory storage. Their adjoint operators, that share the same property, are used in the computation of the derivatives needed in the optimization algorithm and their physical interpretation is also given. This makes it possible to use matrix-free optimization. The optimization problems are solved using an accelerated version of the FBS.

Numerical simulations are then presented where RIRs generated with a modified version of the IM are interpolated using the described algorithms. The KCV method is used to select the regularization parameter and it is shown that this strategy works correctly if enough microphone measurements are used. The different approaches to perform RIR interpolation are then compared. The comparison shows that the novel spatio-temporally sparse representation outperforms the others particularly when the source signal is available. This provides a good RIR interpolation even when only few microphone measurements are available. Finally, the different approaches are applied using actual measured RIRs reaching similar results to the one obtained using simulated data.

ACKNOWLEDGMENT

The authors would like to thank Lorenzo Stella, Andrea Themelis and Panagiotis Patrinos for their suggestions and precious help in the algorithm design and coding.

REFERENCES

[1] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 249–262, 2002.

[2] A. T. Rosell, "Methods of measuring impulse responses in architectural acoustics," Master's thesis, Master Thesis at the Technical University of Denmark (DTU) in Lyngby, 2009.

[3] T. Ajdler, L. Sbaiz, and M. Vetterli, "Dynamic measurement of room impulse responses using a moving microphone," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1636–1645, 2007.

[4] ——, "The plenacoustic function and its sampling," *IEEE Trans. Signal Process.*, vol. 54, no. 10, pp. 3790 –3804, Oct. 2006.

[5] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[6] A. Tarantola, *Inverse problem theory and methods for model parameter estimation*. siam, 2005.

[7] G. P. Nava, Y. Yasuda, Y. Sato, and S. Sakamoto, "On the in situ estimation of surface acoustic impedance in interiors of arbitrary shape by acoustical inverse methods," *Acoustical Science and Technology*, vol. 30, no. 2, pp. 100–109, 2009.

[8] N. Antonello, T. van Waterschoot, M. Moonen, and P. A. Naylor, "Identification of surface acoustic impedances in a reverberant room using the FDTD method," in *Acoustic Signal Enhancement (IWAENC), 2014 14th International Workshop on*. IEEE, 2014, pp. 114–118.

[9] ——, "Evaluation of a numerical method for identifying surface acoustic impedances in a reverberant room," in *Proc. of the10th European Congress and Exposition on Noise Control Engineering*, 2015, pp. 1–6.

[10] K. Kowalczyk and M. van Walstijn, "Room acoustics simulation using 3-D compact explicit FDTD schemes," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 19, no. 1, pp. 34–46, 2011.

[11] T. Betlehem and T. D. Abhayapala, "Theory and design of sound field reproduction in reverberant rooms," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2100–2111, 2005.

[12] R. A. Kennedy, P. Sadeghi, T. D. Abhayapala, and H. M. Jones, "Intrinsic limits of dimensionality and richness in random multipath fields," *IEEE Transactions on signal processing*, vol. 55, no. 6, pp. 2542–2556, 2007.

[13] P. Samarasinghe, T. Abhayapala, M. Poletti, and T. Betlehem, "An efficient parameterization of the room transfer function," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2217–2227, 2015.

[14] B. Bu, T. D. Abhayapala, C.-c. Bao, and W. Zhang, "Parameterization of the three-dimensional room transfer function in horizontal plane," *The Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. EL280–EL286, 2015.

[15] A. Moiola, R. Hiptmair, and I. Perugia, "Vekua theory for the helmholtz operator," *Zeitschrift für angewandte Mathematik und Physik*, vol. 62, no. 5, pp. 779–807, 2011.

[16] R. Mignot, G. Chardon, and L. Daudet, "Low frequency interpolation of room impulse responses using compressed sensing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 205–216, 2014.

[17] W. Jin and W. B. Kleijn, "Theory and design of multizone soundfield reproduction using sparse methods," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2343–2355, 2015.

[18] R. Mignot, L. Daudet, and F. Ollivier, "Room reverberation reconstruction: Interpolation of the early part using compressed sensing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2301–2312, 2013.

[19] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[20] A. Sarkissian, "Method of superposition applied to patch near-field acoustic holography," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 671–678, 2005.

[21] E. Fernandez-Grande, "Sound field reconstruction using a spherical microphone array," *The Journal of the Acoustical Society of America*, vol. 139, no. 3, pp. 1168–1178, 2016.

[22] E. Fernandez-Grande and A. Xenaki, "Compressive sensing with a spherical microphone array," *The Journal of the Acoustical Society of America*, vol. 139, no. 2, pp. EL45–EL49, 2016.

[23] S. Spors, R. Rabenstein, and J. Ahrens, "The theory of wave field synthesis revisited," in *124th AES Convention*, 2008, pp. 17–20.

[24] S. Tervo, J. Ptynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *J. Audio Eng. Soc*, vol. 61, no. 1/2, pp. 17–28, 2013.

[25] P.-A. Gauthier, A. Berry, and W. Woszczyk, "Sound-field reproduction in-room using optimal control techniques: Simulations in the frequency domain," *The Journal of the Acoustical Society of America*, vol. 117, no. 2, pp. 662–678, 2005.

[26] M. Kolundzija, C. Faller, and M. Vetterli, "Reproducing sound fields using MIMO acoustic channel inversion," *Journal of the Audio Engineering Society*, vol. 59, no. 10, pp. 721–734, 2011.

[27] 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.

[28] N. Radmanesh and I. S. Burnett, "Generation of isolated wideband sound fields using a combined two-stage LASSO-LS algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 378–387, 2013.

[29] N. Parikh and S. P. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[30] L. Stella, A. Themelis, and P. Patrinos, "Forward-backward quasi-newton methods for nonsmooth optimization problems," *Computational Optimization and Applications*, vol. 67, no. 3, pp. 443–487, 2017.

[31] A. Themelis, L. Stella, and P. Patrinos, "Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone line-search algorithms," *arXiv preprint arXiv:1606.06256*, 2016.

[32] S. Diamond and S. Boyd, "Matrix-free convex optimization modeling," *arXiv preprint arXiv:1506.00760*, 2015.

[33] J. Folberth and S. Becker, "Efficient adjoint computation for wavelet and convolution operators [lecture notes]," *IEEE Signal Processing Magazine*, vol. 33, no. 6, pp. 135–147, 2016.

[34] M. Costabel, "Time-dependent problems with the boundary integral equation method," *Encyclopedia of Computational Mechanics*, 2004.

[35] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, "Splitting the unit delay," *IEEE Signal Process. Mag.*, vol. 13, no. 1, pp. 30–60, January 1996.

[36] Á. González, "Measurement of areas on a sphere using Fibonacci and latitude–longitude lattices," *Mathematical Geosciences*, vol. 42, no. 1, pp. 49–64, 2010.

[37] R. Marques, C. Bouville, M. Ribardière, L. P. Santos, and K. Bouatouch, "Spherical Fibonacci point sets for illumination integrals," in *Computer Graphics Forum*, vol. 32, no. 8. Wiley Online Library, 2013, pp. 134–143.

[38] L. Stella and N. Antonello. (2016) Proximaloperators.jl. [Online]. Available: http://github.com/kul-forbes/ProximalOperators.jl https://lirias.kuleuven.be/handle/123456789/587243

[39] ——. (2017) RegLS.jl. [Online]. Available: https://lirias.kuleuven.be/handle/123456789/587256 https://lirias.kuleuven.be/handle/123456789/587243

[40] J. F. Claerbout, *Earth soundings analysis: Processing versus inversion*. Blackwell Scientific Publications Cambridge, Massachusetts, USA, 1992, vol. 6.

[41] E. De Sena, N. Antonello, M. Moonen, and T. Van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 774–786, 2015.

[42] M. Vorlander, *Auralization - Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. Springer-Verlag, 2008.

[43] E. De Sena, H. Hacıhabiboğlu, Z. Cvetković, and J. O. Smith, "Efficient synthesis of room acoustics via scattering delay networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1478–1492, 2015.

[44] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Am.*, vol. 80, no. 5, pp. 1527–1529, 1986.

[45] N. Antonello. (2016) Randomized image method (RIM). [Online]. Available: https://github.com/nantonel/RIM.jl https://lirias.kuleuven.be/handle/123456789/488661

[46] ——. (2017) Room impulse response interpolation using a sparse spatio-temporal representation of the sound field software. [Online]. Available: https://lirias.kuleuven.be/handle/123456789/587256

[47] S. Theodoridis, *Machine learning: a Bayesian and optimization perspective*. Academic Press, 2015.

[48] T. Kronvall, F. Elvander, S. I. Adalbjörnsson, and A. Jakobsson, "Multi-pitch estimation via fast group sparse learning," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 1093–1097.

[49] J. Pätynen, S. Tervo, and T. Lokki, "Analysis of concert hall acoustics via visualizations of time-frequency and spatiotemporal responses," *The Journal of the Acoustical Society of America*, vol. 133, no. 2, pp. 842–857, 2013.

[50] J. K. Nielsen, J. R. Jensen, S. H. Jensen, and M. G. Christensen, "The single- and multichannel audio recordings database (SMARD)," in *Int. Workshop Acoustic Signal Enhancement*, Sep. 2014. [Online]. Available: http://www.smard.es.aau.dk/



Niccolò Antonello received his B.Sc. in Electronic Engineering at the Università degli Studi di Padova (Italy) and his M.Sc. in Acoustic Engineering at the Technical University of Denmark (DTU) in 2010 and 2012, respectively. He is currently pursuing the Ph.D. degree at KU Leuven (Belgium) as an Early Stage Researcher in the Marie Curie Initial Training Network "Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS)". His research interests include room acoustic, inverse problems and numerical optimization.



Enzo De Sena (S11, M14) received his B.Sc. degree in 2007 and his M.Sc. degree cum laude in 2009, both from the Universitá degli Studi di Napoli Federico II in telecommunication engineering. In 2013, he received his Ph.D. degree in electronic engineering from King's College London, United Kingdom. Between 2013 and 2016, he was a postdoctoral research fellow at the Katholieke Universiteit Leuven, Belgium. Since September 2016, he has been a lecturer in audio at the Institute of Sound Recording of the University of Surrey, Guildford, United Kingdom. He held visiting positions at the Center for Computer Research in Music and Acoustics at Stanford University, California (2013), in the Signal and Information Processing section at Aalborg University, Denmark (20142015), and in the Speech and Audio Processing Group at Imperial College London, United Kingdom (2016). He is a former Marie Curie fellow, and his current research interests include room acoustics modeling, multichannel audio systems, microphone beamforming, and binaural modeling. He is a Member of the IEEE.



Marc Moonen Marc Moonen (M'94, SM'06, F'07) is a Full Professor at the Electrical Engineering Department of KU Leuven, where he is heading a research team working in the area of numerical algorithms and signal processing for digital communications, wireless communications, DSL and audio signal processing. He received the 1994 KU Leuven Research Council Award, the 1997 Alcatel Bell (Belgium) Award (with Piet Vandaele), the 2004 Alcatel Bell (Belgium) Award (with Raphael Cendrillon), and was a 1997 Laureate of the Belgium Royal Academy of Science. He received journal best paper awards from the IEEE Transactions on Signal Processing (with Geert Leus and with Daniele Giacobello) and from Elsevier Signal Processing (with Simon Doclo). He was chairman of the IEEE Benelux Signal Processing Chapter (1998-2002), a member of the IEEE Signal Processing Society Technical Committee on Signal Processing for Communications, and President of EURASIP (European Association for Signal Processing, 2007-2008 and 2011-2012). He has served as Editor-in-Chief for the EURASIP Journal on Applied Signal Processing (2003-2005), Area Editor for Feature Articles in IEEE Signal Processing Magazine (2012-2014), and has been a member of the editorial board of IEEE Transactions on Circuits and Systems II, IEEE Signal Processing Magazine, Integration-the VLSI Journal, EURASIP Journal on Wireless Communications and Networking, and Signal Processing. He is currently a member of the editorial board of EURASIP Journal on Advances in Signal Processing.



Patrick A. Naylor (M89, SM'07) is a member of academic staff in the Department of Electrical and Electronic Engineering at Imperial College London. He received the BEng degree in Electronic and Electrical Engineering from the University of Sheffield, UK, and the PhD. degree from Imperial College London, UK. His research interests are in the areas of speech, audio and acoustic signal processing. He has worked in particular on adaptive signal processing for dereverberation, blind multichannel system identification and equalization, acoustic echo control, speech quality estimation and classification, single and multi-channel speech enhancement and speech production modelling with particular focus on the analysis of the voice source signal. In addition to his academic research, he enjoys several fruitful links with industry in the UK, USA and in Europe. He is the past-Chair of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, a director of the European Association for Signal Processing (EURASIP) and formerly an associate editor of IEEE Signal Processing Letters and IEEE Transactions on Audio Speech and Language Processing.



Toon van Waterschoot (S'04, M'12) received the MSc (2001) and PhD (2009) degrees in Electrical Engineering, both from KU Leuven, Belgium, where he is currently a tenure-track Assistant Professor. He has previously held teaching and research positions with the Antwerp Maritime Academy, the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT), and the Research Foundation - Flanders (FWO) in Belgium, with Delft University of Technology in The Netherlands, and with the University of Lugano in Switzerland. His research interests are in signal processing, machine learning, and numerical optimization, applied to acoustic signal enhancement, acoustic modeling, audio analysis, and audio reproduction. He has been serving as an Associate Editor for the Journal of the Audio Engineering Society (AES) and for the EURASIP Journal on Audio, Music, and Speech Processing, and as a Guest Editor for Elsevier Signal Processing. He is a Member of the Board of Directors of the European Association for Signal Processing (EURASIP) and a Member of the IEEE Audio and Acoustic Signal Processing Technical Committee (AASP-TC). He was the General Chair of the 60th AES International Conference in Leuven, Belgium (2016), and has been serving on the Organizing Committee of the European Conference on Computational Optimization (EUCCO 2016) and the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2017). He is a member of EURASIP, IEEE, ASA, and AES.