Sundström, David; Elvander, Filip; Jakobsson, Andreas

Estimation of Impulse Responses for a Moving Source Using Optimal Transport Regularization

Published in:
ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

DOI:
10.1109/ICASSP48485.2024.10446838

Published: 19/04/2024

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Sundström, D., Elvander, F., & Jakobsson, A. (2024). Estimation of Impulse Responses for a Moving Source Using Optimal Transport Regularization. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 921-925). Article 10446838 (IEEE International Conference on Acoustics, Speech, and Signal Processing proceedings). IEEE.
https://doi.org/10.1109/ICASSP48485.2024.10446838

Optimal Transport Based Impulse Response Interpolation in the Presence of Calibration Errors

David Sundström, Filip Elvander, and Andreas Jakobsson

***Abstract*—Acoustic impulse responses (IRs) are widely used to model sound propagation between two points in space. Being a point-to-point description, IRs are generally estimated based on input-output pairs for source and sensor positions of interest. Alternatively, the IR at an arbitrary location in space may be constructed based on interpolation techniques, thus alleviating the need of densely sampling the space. The resulting IR interpolation problem is of general interest, e.g., for imaging of subsurface structures based on seismic waves, rendering of audio and radar IRs, as well as for numerous spatial audio applications. A commonly used model represents the acoustic reflections as image sources, often being determined using a sparse reconstruction framework employing spatial dictionaries. However, in the presence of calibration errors, such spatial dictionaries tend to inaccurately represent the actual propagation, limiting the use of these methods in practical applications. Instead of explicitly assuming an image source model, we here introduce a trade-off between minimizing the distance to an image source model and fitting the data by means of a multi-marginal optimal transport problem. The proposed method is evaluated on the early part of real acoustic IRs from the MeshRIR data set, illustrating its preferable performance as compared to state-of-the-art spatial dictionary-based IR interpolation approaches.**

***Index Terms*—Optimal mass transport, impulse response interpolation, Robust time-delay estimation**

I. Introduction

RECENTLY, several spatial audio applications, such as auralization [2], virtual reality [3], spatial active noise control [4]–[7], and the creation of individual sound zones [8], [9] have emerged, which has stimulated the development of spatial modelling of IRs. Such applications all rely on the interpolation of measured IRs from a sensor array to spatial positions where it might not be possible or desirable to place a microphone. While the interpolation problem has attracted substantial attention within the audio community, its relevance extends to various other applications, such as imaging of subsurface structures based on seismic waves [10] and environmental monitoring using synthetic aperture radar [11].

Acoustic room IRs are commonly divided into an early and a late part, where the early part generally consists of clearly distinguishable reflections, whereas the late part models a more diffuse field [12]. For applications such as spatial active noise control and the generation of individual sound zones, the early part of the IR is of main interest as the spatial information content of the audio signal is marginal in the late part [13], [14]. For other applications, also the late part is of importance, and there is a rich literature on low frequency interpolation, commonly utilizing a parameterization of the wave equation [15]–[24]. Multiple approaches for interpolating the early part of IRs have been proposed, commonly exploiting some spatial representation of the IR. A further interesting development is the recent machine learning based IR interpolation approaches (see, e.g., [25]–[28]), which has been reported to show promising results, although these methods typically require using a large number of sensors, in the range of hundreds, to learn the sought parameters. In [29]–[33], the room geometry of the problem is assumed to be known and the impedances of the reflections are estimated as an inverse problem. Regrettaby, the room geometry is rarely known in practice, making the problem ill posed without further priors on the parameters. As an alternative, to allow for an interpolation without assuming prior information of the room geometry, a plane wave assumption has often been used [17], [23], [34], [35]. Although it has been shown that plane waves can represent every solution to the homogeneous wave equation [36], the spherical propagation of waves requires the use of large plane wave dictionaries to be able to achieve accurate approximations. Instead, the spatial representation may be formulated by some variation of the image source[1] method [37], [39], [40]. These methods assume that a reflection can be represented by an image source, such that the delays of the contributions in the IRs corresponds to the distance to where the image source is positioned (see, e.g., [12] for an overview and historical remarks of the development). The identification of image sources from measured IRs is a general problem with applications also within domains such as radar, sonar, and biomedicine. For example, room geometry reconstruction and geometry calibration is investigated using both audio [41]–[45] and radar [46] techniques; for both, the common approach is to identify locations of image sources, which may then be used to reveal reflecting surfaces [41], [47]–[50]. Another application is the localization of objects based on channel estimates, as done for instance by sonar systems [51], [52]. In such systems, the interest is typically to accurately determine the position and shape of an underwater reflector. A similar problem also occurs in many forms of biomedical applications, such as when using EEG to localize brain activity [53].

In the context of IR interpolation, solutions based on the

D. Sundström and A. Jakobsson are with the Centre for Mathematical Sciences, Lund University, Sweden. F. Elvander is with the Dept. of Information and Communications Engineering, Aalto University, Finland. Parts of this work has also been published as [1].

[1]By the term image source model, we refer to the method introduced in [37], where each amplitude-delay pair is associated with an image source position computed based on the positions of the source, receiver, and boundaries. In contrast, by the equivalent source model, we refer to arbitrary source positions, not necessarily being related to the geometry of the room [17], [38].

image source method have also been examined, see e.g., [38], [54]–[56], where the IRs are regularized with sparse priors in the time domain. In practice, estimated IRs are not sparse in the time domain due to the presence of diffuse reflections [57]. To allow for non-sparse IRs, whilst still regularizing the inverse problem, recent attention has shifted to formulations utilizing group sparsity priors [17], [58]. In [17], the room IR interpolation problem is considered for both a time and frequency domain formulation based on equivalent sources, where different types of regularization, namely ridge regression, lasso, and group lasso regularization are compared in the two domains. For the group lasso regularization, a group is defined by a direction of arrival (DOA) and is formed by all of its range entries, which promotes a sparse solution in DOA while still allowing for a linear response of the source and the reflections. The size of the dictionary does however grow rapidly when the delays between the source and sensors are large. In practice, measurements are also corrupted by calibration errors[2], causing a mismatch between the explicit dictionary of the equivalent source positions and the measurements. The calibration errors are especially significant for frequencies where the calibration errors is of the same magnitude as the wavelength, which therefore is a key problem for practical use of previously mentioned applications.

To alleviate this problem, this work aims at allowing for such calibration errors by formulating the interpolation problem using an optimal mass transport formulation. Recently, the concept of optimal mass transport (OMT) has attracted increasing attention as a tool for quantifying the distance between two distributions. Originating from the early work of Kantorovich, the distance between two discrete distributions can be defined by their pointwise distance, where not only the energy of the signal is considered, but also the location of the energy (see [59], [60] for overviews). While originating from the economic community, implementations of OMT distances has recently found applications within the signal processing community in problems such as spectral estimation [61]–[64], localization and sensor fusion [65]–[67], and more (see for example [68]). The growing interest in the area has pushed the development of computationally efficient solvers and the development of formulations allowing for aspects such as the transport between distributions of different total mass [69] and for defining distances between multiple distributions, which has led to the development of multi-marginal OMT [70]. The strength of the OMT formulation lies in its flexibility to design distances based on the location of the mass, making it natural to exploit in problems where both time and space are of relevance. The use of OMT in IR interpolation has recently been proposed in [71], where the problem of interpolating the temporally sparse IRs given their image source representations is studied. The formulation is of relevance for acoustic rendering with known image source positions, where the problem is to linearly interpolate the known image source positions between two impulse responses. In this work, we instead assume that only the impulse responses and the sensor positions are known, and study the problem of interpolating an IR in the presence of calibration errors. To do so, we propose a trade-off between minimizing the distance to an image source model and fitting the data by introducing a multi-marginal OMT distance.

[2]By the term calibration errors, we here refer to (deterministic) model mismatch affecting the delay structure of the IRs, caused by, e.g., errors in the assumed sensor positions or the non-isotropic propagation channels.

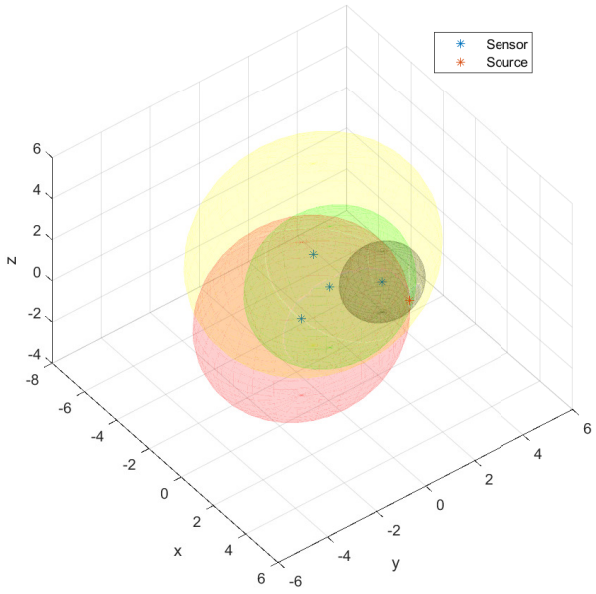


Fig. 1. Illustration of the cost in (6), where $\mathbf{C}_{i_1,\ldots,i_M} = 0$, i.e., the spheres centered around the sensor positions with radii $i_1 c/f_s, \ldots, i_M c/f_s$ have a unique intersection point, here marked as the source position.

The remainder of this paper is organized as follows: in the next section, the signal model is introduced. In Section III, we introduce the multi-marginal OMT formulation, which is extended with unbalanced transportation in Section III-A. An introduction of the joint estimation of the frequency response of the sources and reflections is presented in Section III-B, whereas the selection of the required regularization parameters is discussed in Section III-C. Section III-D introduces the interpolation of fractional delays, followed in Section III-E by practical implementation aspects to reduce the size of the problem, and by details on how to describe the interpolation as a forward problem in Section III-F. Then, in Section IV, the method is evaluated on real audio data, confirming the advantages of the method in a practical scenario. Finally, the work is concluded in Section V.

II. Signal Model

Consider a source at an unknown position $\mathbf{y} \in \mathbb{R}^3$ emitting a signal $\mathbf{z} \in \mathbb{R}^N$ which is recorded by S sensors with known positions $\mathbf{x}_s \in \mathbb{R}^3$, for $s = 1, \ldots, S$. The recorded signal $\mathbf{s}_s \in \mathbb{R}^{N+N_h-1}$ may thus be described as

$$\mathbf{s}_s = \mathbf{h}_s * \mathbf{z} + \boldsymbol{\epsilon}_s, \tag{1}$$

where $\mathbf{h}_s \in \mathbb{R}^{N_h}$ denotes the IR between the points $\mathbf{y}$ and $\mathbf{x}_s$, $*$ the convolution operator, and $\boldsymbol{\epsilon}_s \in \mathbb{R}^{N+N_h-1}$ an additive Gaussian noise with variance σ_ϵ^2. The IR $\mathbf{h}_s$ can be represented

as amplitude-delay pairs by the set $\{(o_{k,s}, \tau_{k,s})\}_k$, where $o_{k,s}$ denotes the amplitude and $\tau_{k,s}$ the delay of the kth filter tap. The geometrical interpretation of the IR under the assumption of perfect specular reflections, $\tilde{\mathbf{h}}_s$, then indicates the presence of a source with amplitude $o_{k,s}$ located on a sphere with radius proportional to $\tau_{k,s}$ centered around the sensor position $\mathbf{x}_s$. As illustrated in Figure 1, when spheres of the taps of at least 4 IRs intersect in a single point, this indicates the presence of a source in the intersection point. An equivalent way of representing the IRs under ideal conditions is thus as amplitude-source position pairs, which is also known as the image source model. However, in the presence of calibration errors, e.g., due to errors in the assumed sensor position or sampling, the spheres will not provide a clear point of intersection, resulting in a mismatch in the amplitude-source position representation. It is worth stressing that, different from the interpolation problem considered in [71], which assumes known image source positions, we here only assume the IRs $\mathbf{h}_s$ and the sensor locations to be known.

A commonly used simplified model of the IRs treats the amplitudes as formed by a sparse representation under the assumption of perfect specular reflections. Regrettably, this model does not accurately represent realistic IRs, as these are generally corrupted by various reflections as well as by the characteristics of the sound source. Here, we instead assume a simplified model for these distortions, modelling these using finite impulse responses. As it is often difficult to distinguish between the distortions due to the source and the various reflections, we will here model these jointly, denoting the joint linear frequency response of an image source a *signature*. Due to directivity of the sources and the corresponding reflections, the signature of each image source may differ. Figure 2 shows how a measured IR, $\mathbf{h}_s$, containing a direct path and 3 reflections, i.e., 4 components, may be decomposed into 4 signatures, $\mathbf{h}_{k,signature}$, and a sparse IR, $\tilde{\mathbf{h}}_s$, such that $\mathbf{h}_s = \sum_k \mathbf{h}_{k,signature} * \tilde{\mathbf{h}}_s$. Given the signatures $\mathbf{h}_{k,signature}$ and their image source positions, the IR may then be interpolated to a position $\hat{\mathbf{x}} \in \mathbb{R}^3$ as a forward problem. In the following, we consider the problem of predicting the IR at an location $\hat{\mathbf{x}} \in \mathbb{R}^3$ given the IRs $\mathbf{h}_s$ and the corresponding sensor positions $\mathbf{x}_s$ in the presence of calibration errors, i.e., allowing for a mismatch between the data and the amplitude-source position representation.

III. Method

We proceed to introduce an OMT formulation for IR interpolation. In order to allow for the considered calibration errors, we strive to define a weaker description of the image source positions. To define this distance to an image source model, we employ a multi-marginal OMT formulation, such that mass is transported in a geometrically consistent manner, encouraging well defined spatial solutions. By doing so, an explicit dictionary of the spatial source distribution does not have to be constructed, instead allowing for a formulation using a dictionary of set of delays.

The classical discrete optimal transport problem is defined between two distributions $\mathbf{\Phi}_1 \in \mathbb{R}^{N_1}$ and $\mathbf{\Phi}_2 \in \mathbb{R}^{N_2}$, where

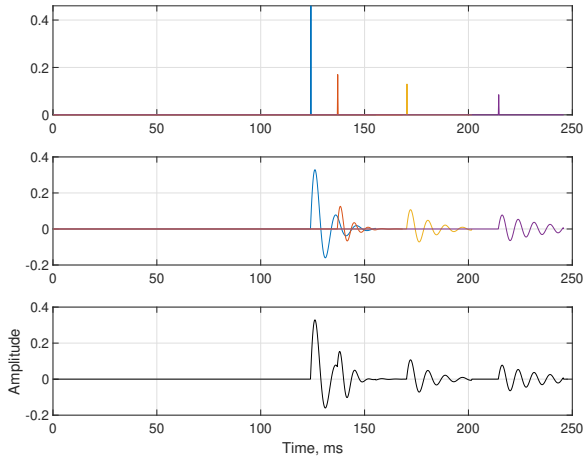


Fig. 2. An illustrative example of how a measured IR can be decomposed into a sparse IR $\tilde{\mathbf{h}}_s$, shown in the top figure, and a set of signatures, $\mathbf{h}_{k,signature}$, shown in the middle figure. The measured IR $\mathbf{h}_s = \sum_k \mathbf{h}_{k,signature} * \tilde{\mathbf{h}}_s$ is illustrated in the bottom figure.

the problem is to move the mass from $\mathbf{\Phi}_1$ to $\mathbf{\Phi}_2$. The cost of moving the mass from index i_1 to index i_2 is defined by the cost $\mathbf{C}_{i_1,i_2}$, where $\mathbf{C} \in \mathbb{R}^{N_1 \times N_2}$ is the cost matrix. With this, the Kantorovich problem of OMT may be formulated as

$$\begin{aligned} \underset{\mathbf{M} \in \mathbb{R}_+^{N_1 \times N_2}}{\text{minimize}} \quad & \sum_{i_1,i_2} \mathbf{C}_{i_1,i_2} \mathbf{M}_{i_1,i_2} \\ \text{s.t.} \quad & \mathbf{M}\mathbf{1}_{N_2} = \mathbf{\Phi}_1 \ , \ \mathbf{M}^T \mathbf{1}_{N_1} = \mathbf{\Phi}_2, \end{aligned} \tag{2}$$

where $\mathbf{M} \in \mathbb{R}_+^{N_1 \times N_2}$ is the transport plan between $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$, and $\mathbf{1}_{N_1}$ and $\mathbf{1}_{N_1}$ are vectors of ones. Here, $\mathbf{M}$ may be interpreted as a distribution on the product space $\mathbf{R}^{N_1} \times \mathbf{R}^{N_2} = \mathbf{R}^{N_1 \times N_2}$ having $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ as its marginals, and for which the structure is determined by the cost $\mathbf{C}$. We will here consider the transport problem between IRs at two different sensor positions, i.e., $\mathbf{\Phi}_1 = \mathbf{h}_1$ and $\mathbf{\Phi}_2 = \mathbf{h}_2$, defining the cost $\mathbf{C}$ to establish a geometrically meaningful association of delays in the transport plan $\mathbf{M}$. Although the classical transport problem in (2) defines the distance between two distributions, it can be extended to S distributions via a multi-marginal optimal transport formulation where $\mathbf{M}$ and $\mathbf{C}$ are formed as tensors. Although the intuition in terms of transportation of mass is more involved for the multi-marginal setting, the interpretation of the transport tensor $\mathbf{M}$ in terms of joint distributions still holds such that the multi-marginal transport problem considers computing a joint distribution between S marginal distributions with the cost tensor $\mathbf{C}$ determining the structure of the joint distribution. We note that the multi-marginal transport problem allows for efficiently modeling collections of coupled pair-wise transport problems (see for example [65]), and its general formulation does allow for inducing more precise structure than what is possible in a pair-wise setting. Below, it is shown that the multi-marginal formulation is necessary to exploit all of the geometrical structure inherited from the image source method. In particular, the multi-marginal formulation will allow us to

model the full joint delay structure of a set of IRs using a small number of image sources.

To introduce our formulation, consider S sensors measuring a set of sparse and positive IRs[3], $\mathbf{h}_s \in \mathbb{R}_+^{N_h}$. Clearly, this is a strong assumption. It is introduced here in order to build the intuition for the proposed method, but is then relaxed in Section III-B. Forming a multi-marginal version of Kantorovich's transport formulation allows the association, or the transport plan, of the mass between the IRs to be given as the solution to

$$\begin{aligned} \underset{\mathbf{M}\in\mathbb{R}_+^{N_h^S}}{\text{minimize}} \quad & \langle \mathbf{C}, \mathbf{M} \rangle \\ \text{s.t.} \quad & \mathcal{P}_s(\mathbf{M}) = \mathbf{h}_s,\ s = 1, \ldots, S, \end{aligned} \tag{3}$$

where

$$\langle \mathbf{C}, \mathbf{M} \rangle := \sum_{i_1,\ldots,i_S} \mathbf{C}_{i_1,\ldots,i_S} \mathbf{M}_{i_1,\ldots,i_S}, \tag{4}$$

with $\mathbf{M}$ denoting the S-fold transport plan tensor, $\mathbf{C}$ the corresponding cost tensor, and $\mathcal{P}_s$ the projection on the sth marginal, defined as

$$\mathcal{P}_s(\mathbf{M})_j = \sum_{i_1,\ldots,i_{s-1},i_{s+1},\ldots,i_S} \mathbf{M}_{i_1,\ldots,i_{s-1},j,i_{s+1},\ldots,i_M}. \tag{5}$$

Here, the shorthand

$$\mathbb{R}^{N_h^S} := \prod_{s=1}^{S} \mathbb{R}^{N_h} = \mathbb{R}^{N_h} \times \mathbb{R}^{N_h} \times \ldots \times \mathbb{R}^{N_h}$$

is used for the iterated Cartesian product. As an example, in an ideal free-field propagation scenario, in which each IR, $\mathbf{h}_s$, has only one non-zero element at index, say, k_s, the transport tensor $\mathbf{M}$ has a single non-zero element $\mathbf{M}_{k_1,\ldots,k_S}$. This element then correspond to the direct path components with delays $k_1 c/f_s, \ldots, k_S c/f_s$. It is worth noting that it is here implicitly assumed that the sensors are close enough in space such that the amplitude of each image source is the same in each IR. The effect of this assumption may be quantified by considering a point source and two sensors located at the distances r_1 and r_2 from the source position. Due to spherical propagation, the pressure p_0 by the source decays in amplitude proportional to the distance, $p(r) \propto p_0/r$, such that the error is given by $p(r_2) - p(r_1) \propto p_0(r_1 - r_2)/r_2 r_1$. For example, with $r_1 = 3$ and $r_2 = 3.5$, which could be a typical setup for spatial audio applications (see Section IV), the error due to the assumption of equal amplitudes introduces a relative error of less than 3%. We note that a propagation model could also be included in the model, introducing a scaling factor of the projection in (5), which would be of interest in applications with non-homogeneous propagation such as sonar or EEG.

The solution of (3) is determined by the cost tensor $\mathbf{C}$. Here, we seek to define the cost to promote transportation of mass between taps of the IRs that correspond to positions in space. As motivated in Section II, the positions are not well defined in the presence of calibration errors. However, it is still possible to define a position in terms of the mean squared error, i.e., the point $\mathbf{y}$ that minimizes the mean squared error to each sphere defined by the delays $\tau_{i_s,s}$. In this manner, the cost of transporting mass between the indices $i_1, \ldots, i_S$ may be defined by the corresponding deviation of the source position $\mathbf{y}$ such that

$$C_{i_1,\ldots,i_S} = \min_{\mathbf{y}\in\mathbb{R}^3} \sum_{s=1}^{S} \Big| ||\mathbf{x}_s - \mathbf{y}||_2 - \tau_{i_s,s} c \Big|^2, \tag{6}$$

where c is the propagation speed. The problem in (6) is a so-called trilateration problem; an illustration of the setup is shown in Figure 1. Assuming Gaussian noise, the resulting optimization is generally non-linear, non-convex, and multi-modal. Fortunately, a computationally efficient approximate solution have been proposed (see, e.g., [72]). Given the definition in (6), elements of the transport tensor $\mathbf{M}$ that correspond to delays with smaller intersection errors of the spheres will yield lower costs in (3). It is worth noting that the cost in (6) is agnostic of the IRs, $\mathbf{h}_s$, being solely defined by the sensor positions. In the following, the geometrical transport concept is extended with probabilistic measurement models and implementation considerations to handle identification of image sources based on realistic estimated IRs.

A. Robustness for noisy impulse responses

The formulation in (3) requires the total transported mass to remain constant due to the hard marginal constraints. However, IRs that are estimated under non-ideal conditions implies that the total mass of the resulting IRs may vary. Here, the estimated IRs, $\mathbf{h}_s$, are assumed to be corrupted by Gaussian noise. Several methods have been proposed to allow also for unbalanced optimal transport depending on the error model of the marginals. To allow for the Gaussian error model, the marginal constraints may be relaxed, such that

$$\underset{\mathbf{M}\in\mathbb{R}_+^{N_h^S}}{\text{minimize}} \quad \langle \mathbf{C}, \mathbf{M} \rangle + \lambda \sum_{s=1}^{S} ||\mathcal{P}_s(\mathbf{M}) - \mathbf{h}_s||_2^2. \tag{7}$$

In this form, mass may both be introduced and discarded from the estimated IRs. If an image source is observed in all but one IR, it may then still be possible to identify the source. However, the relaxed problem in (7) is only of relevance if the parameter λ, which controls the trade-off between transporting and discarding mass, can be set appropriately. We further discuss how this may be solved below in Section III-C.

B. Joint signature estimation

The above formulations are designed to find the transport plan of idealized sparse IRs. To allow for the unknown signatures introduced in Section II, we here consider joint estimation of the signatures and estimation of transport plan. Without losing the intuition of the optimal transport problem, the transportation is then instead defined to transport the energy of the full signatures defined by $||\mathbf{h}_{k,signature}||_2^2$.

In the interest of notational brevity, we here assume the sensor array to be small enough, such that each sensor may be considered to measure approximately the same signature for each source. It is further worth noting that the typically

[3]For notational simplicity, but without loss of generality for the discussion or the derived method, we let all IRs have the same length N_h.

sparse structure of IRs will then translate to similar sparse priors on the resulting signatures, which are here modelled to also allow for non-sparse responses of each image source. In order to do so, each element in (7) is extended with an IR of length N_{sign}, such that the mass $\mathbf{M} \in \mathbb{R}^{N_h^S}$ constitutes the energy of the signatures $\mathbf{U} \in \mathbb{R}^{N_h^S \times N_{sign}}$, here modeled as

$$\mathbf{M} = \mathcal{T}_2(\mathbf{U}), \tag{8}$$

where $\mathcal{T}_p : \mathbb{R}^{N_h^S \times N_{sign}} \to \mathbb{R}^{N_h^S}$ and is defined such that $p = 1$ corresponds to the ℓ_2 norm and $p = 2$ to the squared ℓ_2 norm over the signature dimension of each entry, defined as

$$\mathcal{T}_p(\mathbf{U})_{i_1,\ldots,i_S} = \left(\sum_{i_{sign}}^{N_{sign}} \mathbf{U}^2_{i_1,\ldots,i_S,i_{sign}} \right)^{p/2}. \tag{9}$$

Incorporating the sparse signature in (7) allows the problem to be extended to

$$\begin{aligned} \underset{\mathbf{U} \in \mathcal{R}^{N^S \times N_{sign}}}{\text{minimize}} \; & \langle \mathbf{C}, \mathcal{T}_2(\mathbf{U}) \rangle + \lambda_2 \sum_{i_1,\ldots,i_S} \mathcal{T}_1(\mathbf{U})_{i_1,\ldots,i_S} \\ & + \lambda_1 \sum_{s=1}^{S} ||\mathcal{P}_s^{sign}(\mathbf{U}) - \mathbf{h}_s||_2^2, \end{aligned} \tag{10}$$

where the projection of the signatures on the marginal s at index j is defined as

$$\mathcal{P}_s^{sign}(\mathbf{U})_j = \sum_{a,b \in \mathbb{A}_j} \sum_{\mathcal{I}_s} \mathbf{U}_{i_1,\ldots,i_{s-1},a,i_{s+1},\ldots,i_M,b}, \tag{11}$$

where

$$\mathcal{I}_s = \{(i_1, \ldots, i_{s-1}, i_{s+1}, \ldots, i_S); i_k \in [1, 2, \ldots, N_h]\}$$

is the set of all indices in $\mathbf{U}$ for a given index a in the sth dimension and a given index b of the signature. The set $\mathcal{A}_j = \{(a, b) \in [1, 2, \ldots, N_h] \times [1, 2, \ldots, N_{sign}]; a + b = j\}$ defines the set of every index a of an IR and index b of the signature that corresponds to a delay j, i.e., such that $a+b = j$. Although the definition of $\mathcal{P}_s^{sign}(\cdot)_j$ now appears to be more involved, the principle is the same as for $\mathcal{P}_s(\cdot)_j$ in (5), i.e., assigning the mass of the transport tensor to the corresponding delay in the IR $\mathbf{h}_s$, although now each set of delays is endowed with a signature.

It may be noted that an important aspect gained from the use of the operator $\mathcal{T}_2(\cdot)$ is that it allows for both positive and negative amplitudes of the estimated IRs. Even though ideal IRs are positive under the image source model, estimated IRs are in practice containing both positive and negative terms due to estimation errors and the frequency response of the source and the reflections. In order to be applicable to such measurements, the problem (9) models transport of the energy of the IRs, thus allowing for negative amplitudes. A similar geometrical intuition of the transportation term in (10) still holds, but where the energy of the signature of each image source is transported instead of each amplitude.

Finally, one may note similarities between the modified optimal transport formulation in (10) and classical estimation methods from a signal perspective. Due to linearity of the observation operator in (11), (10) may be interpreted from a Bayesian perspective with a Gaussian noise model of the observed IRs and group sparse as well as weighted Gaussian priors on the model parameters. Although both the weighting of the Gaussian prior on the model, i.e., the cost tensor $\mathbf{C}$, and the linear observation operator in (11), follows naturally from the optimal transport perspective, we will make use of the Bayesian estimation perspective in the following section.

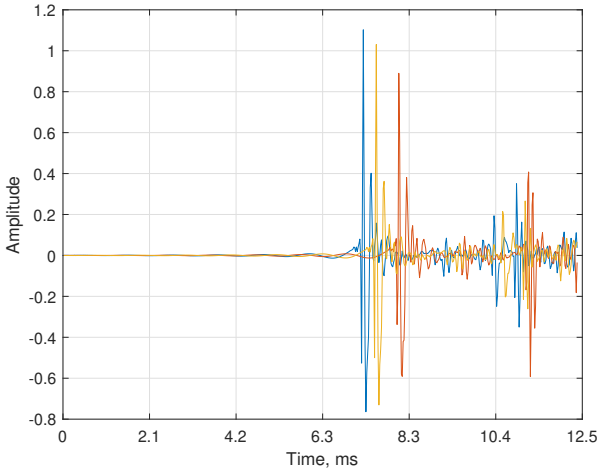


Fig. 3. Illustration of a typical IRs from the MeshRIR dataset [73], where the early part of three IRs are extracted and high-pass filtered with a cut-off frequency of 1000 Hz.

C. Selection of regularization parameters

The trade-off between transporting between IRs and discarding mass is controlled by the regularization parameter λ in (7). This regularization makes the model robust for errors such as sampling mismatches and uncertainties in the sensor positions. We here aim to formulate the trade-off in (7) in a physically meaningful manner, such that λ may be determined based on approximations of these errors. This allows for the selection of the regularization parameters λ_1 and λ_2 in (10) to be determined based on the sensor positions and the noise variance of the data.

Consider assumed sensor positions $\hat{\mathbf{x}}_s$ located within a ball of radius ε_x from the true sensor positions $\mathbf{x}_s$, and time of arrival (TOA) measurements $\hat{\tau}_s$ satisfying $|\hat{\tau}_s - \tau_s| \leq \varepsilon_\tau$, where τ_s denotes the true TOA. Transportation of mass is then only possible if the cost of transportation is not larger than the worst-case cost under this error model for some source position $\mathbf{y} \in \mathbb{R}^3$, i.e.,

$$\begin{aligned} J_{max}(\mathbf{y}) = & \max_{\tilde{\mathbf{x}}_s, \tilde{\tau}_s s=1,\ldots,S} \sum_{s=1}^{S} |||\tilde{\mathbf{x}}_s - \mathbf{y}||_2 - \tilde{\tau}_s c|^p \\ & \text{s.t. } ||\tilde{\mathbf{x}}_s - \mathbf{x}_s||_2 \leq \varepsilon_x \\ & \quad |\tilde{\tau}_s - \tau_s| \leq \varepsilon_\tau \quad s = 1, \ldots, S. \end{aligned} \tag{12}$$

An upper bound of J_{max}, independent of the source position $\mathbf{y}$, is summarized in the following proposition.

Proposition 1. *Consider J_{max} in (12). For any $\mathbf{y} \in \mathbb{R}^3$, it holds that*

$$J_{max}(\mathbf{y}) \leq S(\varepsilon_x + \varepsilon_\tau)^p. \tag{13}$$

Proof. See Appendix A. □

Selecting the regularization parameter to coincide with this upper bound, i.e., such that

$$\lambda = S(\varepsilon_x + \varepsilon_\tau)^p, \tag{14}$$

all feasible transports under this error model are beneficial compared to discarding the mass. In practice, when a sensor array is calibrated using trilateration or multilateration methods, as for example in [72], an estimate of ε_x is available from the mismatch between the measured delays and the estimated position. Furthermore, we let the uncertainties in the TOAs due to sampling determine ε_τ, such that $\varepsilon_\tau = 1/2f_s$. Considering the signature formulation with its group sparse constrains, as defined in (10), the cost function can be seen to consist of one term modelling the room geometry and two terms modelling the IR single-channel data. Similar reasoning as above is therefore valid considering the trade-off between maintaining a geometrically feasible transport plan and modelling the data. The remaining problem of setting the trade-off between the two latter terms in (10), i.e. between the group-sparsity and fit of data, may thus be thought of in terms of setting a parameter ρ to weigh the sum of these two terms, i.e.,

$$\sum_{s=1}^{S} ||\mathcal{P}_s^{sign}(\mathbf{U}) - \mathbf{h}_s||_2^2 + \rho \sum_{i_1,\ldots,i_S} \mathcal{T}_1(\mathbf{U})_{i_1,\ldots,i_S}. \tag{15}$$

It should be noted that this formulation coincides with the well-studied group-sparse regularized least-squares problem. Thus, we will here proceed with the heuristic approach of setting ρ based on the fraction of the largest value that is shrunk to zero, and validate that this is an appropriate choice in Section IV. In the following, it is assumed that the estimated IRs are corrupted by additive Gaussian noise such that $\mathbf{h}_s = \bar{\mathbf{h}}_s + \boldsymbol{\epsilon}_s$, where the noise variance, σ_ϵ^2, may be empirically estimated using the latter part of the IR, $\mathbf{h}_s$, that may be considered to not contain any of the early reflections. A confidence bound of the noise with confidence level α may thus be assumed given by $\sigma_\epsilon \beta_{\alpha/2}$, where $\beta_{\alpha/2}$ denotes the corresponding quantile of the standard Gaussian distribution, such that an estimate of the fraction between the maximum amplitude in the data and the noise floor is $\sigma_\epsilon \beta_{\alpha/2} / \max_{s,n}(\mathbf{h}_s)$, where $\max_{s,n}(\cdot)$ denotes the maximum value of every sample and sensor. Correcting for the group length N_{sign} and the number of sensors S results in

$$\rho = \frac{\sigma_\epsilon \beta_{\alpha/2} N_{sign}}{\max_{s,n}(\mathbf{h}_s) S} \delta, \tag{16}$$

where a confidence value of $\alpha = 0.05$ is used here, and the scaling parameter δ is introduced to allow for a local cross validation with the available data, reminiscent to the parameter selection in [17]. Note that although the formulation in (10) includes two regularization parameters, the cross validation only considers a search over a local region for the parameter δ, which is here performed using 5 values of δ in the range

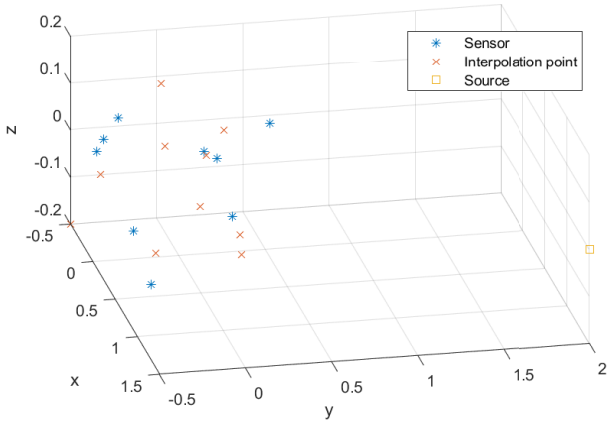


Fig. 4. Illustration of one realization of the data geometry with 10 interpolation points and 13 sensors.

10^{-2} to 10^2. To translate the parameters λ and ρ from (14) and (16) to the parameters in (10), λ_1 and λ_2 are defined as

$$\lambda_1 = \lambda \frac{1}{\rho + 1}, \tag{17}$$

and

$$\lambda_2 = \lambda \frac{\rho}{\rho + 1}, \tag{18}$$

respectively.

D. Interpolating off-grid delays

It is worth noting that the formulation in (10) allows for continuously located sources in space, whereas the estimated IRs, $\mathbf{h}_s$, are sampled on a discrete grid. The delays of the sources are thus likely to be off-grid delays with respect to the IRs. To allow for transport of off-grid delays, errors due to sampling may be mitigated. Also, higher resolution in time implies accurate localization which constitutes the basis of this approach. We further note that the formulation in (10) can be extended to cope with a modified sampling frequency. The off-grid delays may then be incorporated by introducing sinc interpolation assuming narrow-band IRs, followed by a masking procedure, both which can be formulated as linear filters determined by an upsampling factor I_{up} (see [74] for details). Each signature is then initially upsampled by a linear filter to the sampling frequency $f_s I_{up}$, after which the masking is introduced as a linear filter to align the upsampled signatures to the original grid at sampling frequency f_s. In this way, the signatures can be time-delayed with respect to the grid sampled at $f_s I_{up}$, but where the problem in (10) may still be solved for IRs in the original sampling frequency f_s.

E. Pruning the domain of the problem

It is also worth noting that the problem in (10) is numerically intractable, since it grows with the length of the IRs, such that the number of variables in $\mathbf{U}$ is N_h^M, which for $N_h = 500$ and $M = 15$ is roughly 10^{40}. To make the problem tractable and such that the complexity grows with the complexity of the geometry of the environment rather than the length of the IRs,

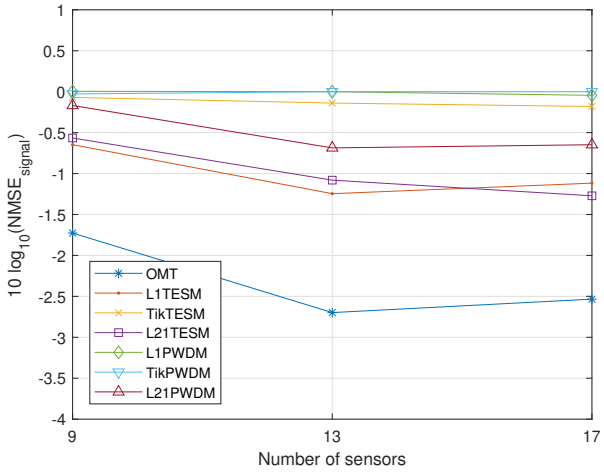


Fig. 5. The figure shows the estimated NMSE$_{signal}$ as a function of number of sensors.

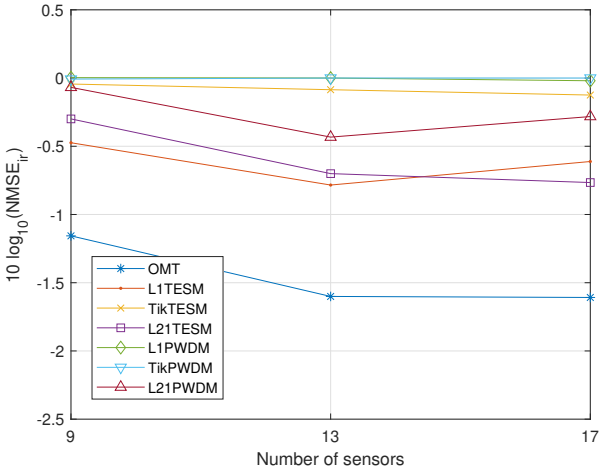


Fig. 6. The figure shows NMSE$_{ir}$ as a function of number of used sensors.

some practical considerations are implemented to propose a set of candidate delays. In the perfect image source model, an IR is sparse with positive peaks corresponding to the delays of the image sources. In the formulation in (10), the image source method is generalized to allow for calibration errors and signatures. The signatures model the response of reflections as well as the response of the sources. It is therefore assumed that each signature attains a maximum which corresponds to the delay to the image source. Therefore, the set of candidate indices may be reduced to the indices that constitute local maximum of the estimated IRs. The local maximum is here determined after up-sampling with the factor I_{up} of the IR, in order to allow for fractional delays. Furthermore, since an estimate of the noise floor usually is available for the early part of the IR, as discussed in Section III-C, only peaks of amplitudes larger than $\lambda_{0.025}\sigma_\epsilon$, i.e., a 95% confidence interval for the noise, are considered. By considering the set of all delays corresponding to a local maximum of the estimated IRs, the number of combination can be further reduced due to Proposition 1. It is known *a priori* that, by setting the trade off between the transport term and the data model terms in (10) as proposed in (14), the set of delays of a cost above the bound in Proposition 1 is outside of the allowed error model. Therefore, these set of delays may also be excluded from the problem, which, along with the other pruning aspects, makes the problem presented in Section IV small enough to be solved using general convex solvers, such as CVX [75]. However, the solution may be even further simplified, by pruning the dictionary size based on the relative amplitudes of the local maximum. It is expected that all of the local maximum should be of about the same magnitude, such that the variance of the amplitudes corresponding to each index $\{i_1, \ldots, i_S\}$ should be small. Therefore, for a fixed i_s, only the, say, 100 elements of smallest variance must be kept in the dictionary. The mentioned pruning aspects is a direct consequence of the proposed formulation constructed by a dictionary of delays, rather than a dictionary of image sources. Although a dictionary of delays generalizes a dictionary of image sources, a dictionary of delays is in the same domain as the observed data, which thus allows for the pruning of the problem. As an example, to produce the results in Figure 5 and 6 takes about two hours using the proposed method on a modern PC, but about two days for the L1TESM method.

F. Interpolating the impulse response

By construction, each signature obtained as a solution to (10) may be localized to provide a representation of the IR similar to the one formed by the image source method. From the spatial representation, a range of problems may thus be addressed, such as the reconstruction of the room geometry, dual-path suppression, or interpolation of the IR (as considered below). The positions of the signatures, $\mathbf{y}^{eq}_{i_1,\ldots,i_M}$, are estimated as the argument minimizing the cost in (6), i.e.,

$$\mathbf{y}^{eq}_{i_1,\ldots,i_S} = \arg\min_{\mathbf{y}} \sum_{s=1}^{S} |||\mathbf{x}_s - \mathbf{y}||_2 - \tau_{i_k,s}c|^2. \qquad (19)$$

Note that the position is already computed in the construction of the cost of the transport problem in (6) and does thus not need to be recomputed. To reconstruct the interpolated IR, $\mathbf{h}_{inter}$, at a position $\mathbf{y}_{inter}$ in the neighborhood of the sensor array, the mass may be delayed to represent the propagation time to the new location. The IR is then estimated as the sum of the elements in $\mathbf{U}$, where each element $\mathbf{U}_{i_1,\ldots,i_S,i_{sign}}$ contributes to the element

$$\mathbf{h}_{inter}\left[||\mathbf{y}^{eq}_{i_1,\ldots,i_S} - \mathbf{y}_{inter}||\frac{f_s}{c} + i_{sign} - 1\right]. \qquad (20)$$

IV. Numerical Results

In order to evaluate the performance of the proposed method, we use both simulated and measured data from subset S1-M3969 of MeshRIR [73]. The simulated data is obtained using the image source method as implemented in [76], with a sampling frequency of 14 kHz, the reflection coefficients for the walls set to 0.5, and using 9261 virtual sources. The source is positioned at $[5.5, 3.5, 2.0]$ in a box-shaped room of

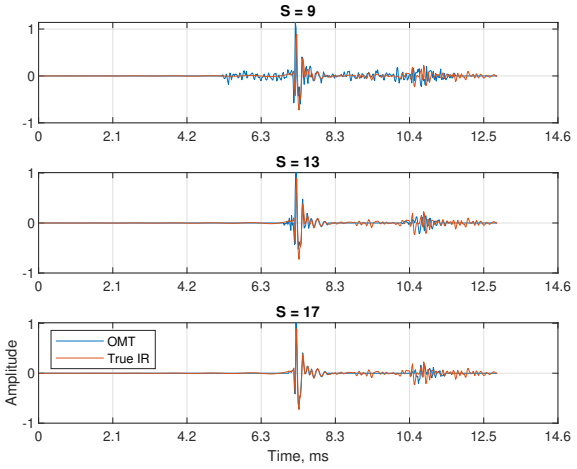


Fig. 7. Example of an interpolated IR formed using by 9, 13, and 17 sensors, respectively.

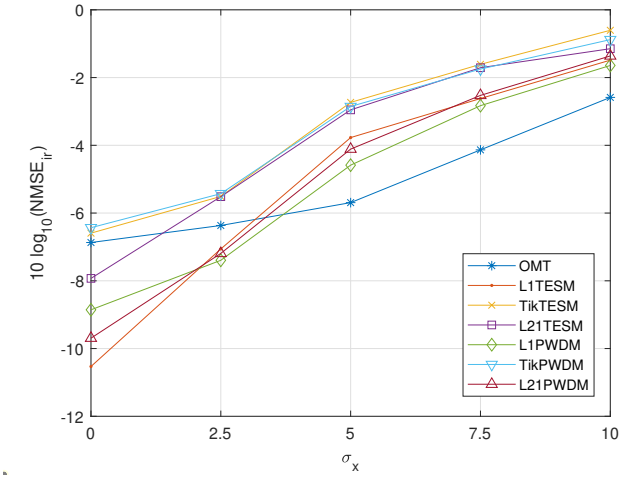


Fig. 8. The figure shows the estimated NMSE_{ir} as a function of the standard deviation for the calibration errors in terms of displacement of the sensor positions for simulated data.

dimensions $11 \times 7 \times 4\text{m}$, with the sensors uniformly sampled in a cube-shaped region centered at $[6.8, 4.2, 1.5]$ with sides of length 0.5m. Furthermore, white Gaussian noise with a standard deviation of 0.005 was added to the simulated IRs. The real data contains estimated IRs sampled at $f_s = 48$ kHz from a single source position to 3969 equally spaced microphones within a measurement region of $1 \times 1 \times 0.4$ m. Since the main contribution of this work is to interpolate the challenging mid and high frequency components of the early part of an IR in the presence of calibration errors, the real IRs are high-pass filtered with a cut-off frequency of 1000 Hz to omit the low frequency dynamics. For low frequency interpolation, which includes aspects such resonant frequencies due to room dimensions, we refer to works with this focus, such as [15]–[24]. Two metrics are presented, averaged over D validation points and F reshuffles. For each reshuffle, S sensors and D validation sensors are randomly chosen from the data set. Figure 4 illustrates a typical realization. Here, we measure the performance of the reconstruction of a signal filtered through the interpolated IR, $\hat{\mathbf{h}}_{p,f}$, using the normalized mean square error (NMSE), defined as

$$\text{NMSE}_{signal} = \frac{1}{DF} \sum_{d=1}^{D} \sum_{f=1}^{F} \frac{||\hat{\mathbf{h}}_{d,f} * \mathbf{z} - \mathbf{h}_{d,f} * \mathbf{z}||_2^2}{||\mathbf{h}_{d,f} * \mathbf{z}||_2^2}, \quad (21)$$

where $\mathbf{h}_{d,f}$ denotes the measured IR from the data set and $\mathbf{z}$ a band-pass filtered signal. Furthermore, to measure the performance of the IR interpolation, define

$$\text{NMSE}_{ir} = \frac{1}{DF} \sum_{d=1}^{D} \sum_{f=1}^{F} \frac{||\hat{\mathbf{h}}_{d,f} * \mathbf{w} - \mathbf{h}_{d,f} * \mathbf{w}||_2^2}{||\mathbf{h}_{d,f} * \mathbf{w}||_2^2}, \quad (22)$$

where $\mathbf{w}$ denotes a low-pass filter. The low-pass filter is introduced since a small deviation of a peak in time would cause a significant contribution to the NMSE_{ir} without any smearing.

In the following, we compare the performance of the proposed method with that of L1TESM, TikTESM, L21TESM, L1PWDM, TikPWDM, and L21PWDM (see [17] for details on these methods, including code for each method), where TESM indicates a method that uses the time domain equivalent source model, whereas PWDM indicates a method that uses the plane wave decomposition model. The used regularization parameter is set using cross-validation, as described in [17]. It is worth noting that the computational complexity of the methods in [17] are very high, necessitating limitations in the numer of used plane wave directions, the assumed lengths of the IRs, the number of sensors, and the number of used cross-validations, in order to make the computationally tractable. This is in particular important for measurements made with a high sampling frequency. Therefore, 10 logarithmically decreasing values for the regularization parameter is evaluated from 1 to 10^{-6}, as described in [17], using 500 equivalent source positions as well as plane wave directions. Unless otherwise specified, the OMT method is evaluated using an up-sampling factor of $I_{up} = 5$, as described in Section III-D, with the signature lengths being $N_{sign} = 40$ for the real data and $N_{sign} = 1$ for the simulated data (this may, in practice, be determined from the first peak where the acoustic IR is commonly separated from the contributions of the reflections). The MeshRIR is calibrated with laboratory equipment such that the sensor position error is assumed to be $\varepsilon_x = 0.01$ m. Moreover, the dictionary of candidate delays is pruned as described in Section III-E.

Initially, simulated data is used to examine the robustness to calibration errors in the sensor positions. To do so, we perturb the sensor positions with a normal distributed error for five different standard deviations, $\sigma_{\mathbf{x}}$, in the range 0 to 10 cm. A sampling frequency 14 kHz is chosen to be able run the methods in [17] on the first 50 ms of the IRs, corresponding to the early part, and with 500 plane wave directions, with D, S, and F set to 10, 30, and 10, respectively. The resulting NMSE_{ir} with the cut-off frequency of the low-pass filter $\mathbf{w}$ in (22) set to 1000 Hz is illustrated in Figure 8. Note that the data is simulated using the image source method without any low frequency considerations, such that a low-pass filter

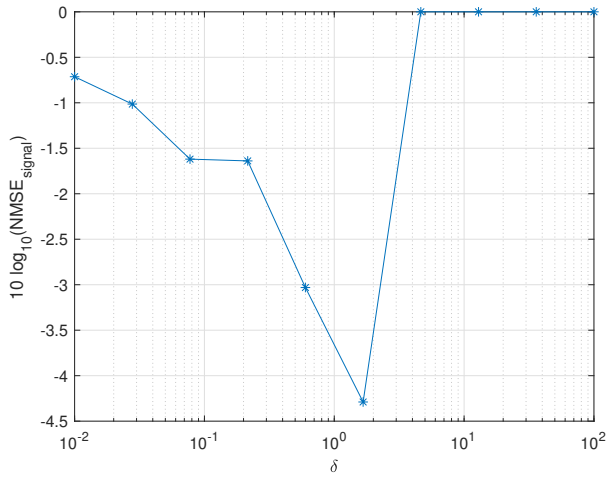


Fig. 9. The figure illustrates the proposed method's selection of the group sparsity regularization parameter ρ, where ρ is scaled with different values of δ such that $\delta = 1$ is equivalent to the proposed selection of ρ.

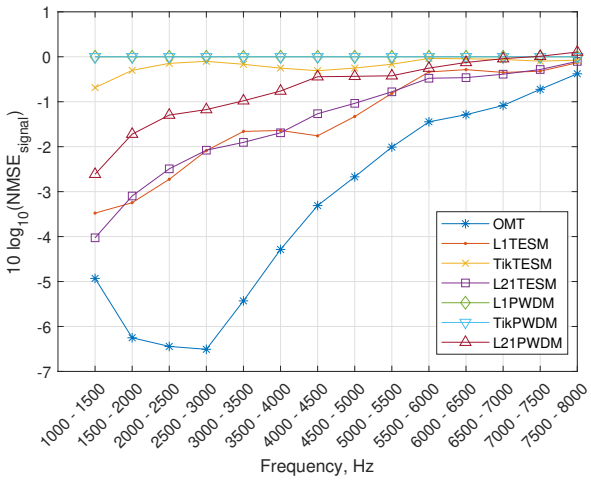


Fig. 10. The figure shows the estimated NMSE of band limited input signals using 13 sensors. The gain in performance of the proposed method is especially prominent for higher frequencies above 2 kHz.

with cut-off frequency 1000 Hz is sufficient for isolating the properties of the early part. Although L1TESM, L21TESM, L1PWDM, and L21PWDM have a lower NMSE_{ir} than the proposed method, denoted OMT, for a perfectly calibrated setting, the OMT can, as expected, be seen to be more robust when introducing calibration errors.

The methods are then examined on the real dataset. Due to the high sampling rate of the IRs, the computational complexity of the methods in [17] is very high. Therefore, the methods are only evaluated for the first 12.5 ms of the IRs, containing only the first 2 to 4 reflections, and with S and F set to 5 and 10, respectively. Below, we also separately evaluate the proposed method on the first 35 ms of the IRs in order to confirm that the obtained results generalizes to a longer part of the early reflections of the IRs. Furthermore, since the real IRs contain various reflections, each such reflection will have a low-passed filtered structure. To mimic this, we use a low-pass filter $\mathbf{w}$ in (22) with cutoff frequency 8 kHz, and a broad band input signal $\mathbf{z}$ in (21), which is band-pass filtered in the range 1 to 5 kHz, to emulate the frequency range of speech.

Initially examining how the number of used sensors affect the performance, Figures 5 and 6 show the NMSE_{signal} and NMSE_{ir} when using 9, 13, and 17 sensors, respectively. As can be seen in the figures, the proposed method consistently outperforms the benchmark methods throughout the interval. We note that the reason for the high NMSE of the benchmark methods could be due to both the calibration errors in the dataset and to the sampling rate, which is 6 time higher than what was used in [17] as well as 3.4 times higher than in the simulation above, making the dictionary immense. The pruning of the proposed solution, as described in Section III-E, which follows directly from formulating the dictionary in the same domain as the data, omits this aspect of the proposed method. An example of the resulting interpolated IR is illustrated in Figure 7, showing the effects of including more sensor measurements.

Next, in order to evaluate the selection of the regularization parameter, described in Section III-C, the NMSE_{signal} is evaluated for different multiplicative scalings of ρ for an interpolation using 13 sensors. The grid of scaling values, δ, is formed over 10 logarithmicly spaced values between 10^{-2} and 10^{2}. As seen in Figure 9, the proposed method of setting the parameter ρ yields a low NMSE_{signal}, with all of the amplitudes in the dictionary being set to zero, yielding an NMSE_{signal} of 0 dB, when δ is greater than about 5.

Figure 10 illustrates how the performance is affected by the frequency content of the signal. Here, the input signal consists of Gaussian noise bandpass filtered in intervals of 500 Hz from 1000 to 8000 Hz, using 13 sensors. As can be seen in the figure, the proposed method offers a clear improvement as compared to the other approaches, most prominently for frequencies in the interval 1500 to 3000 Hz. As can be seen in the figure, for the low frequency signals, the L1TESM and L21TESM perform similar to the proposed method. However, we note that the proposed method is not designed to handle low frequency acoustic components since the signatures are only long enough to capture the frequency response of a single reflection. Acoustic low frequency effects such as room modes have in general longer decay times than the response of a single reflection, which the methods in [17] can therefore handle more efficiently. Finally, Figure 11 illustrates the performance of the proposed method when instead using the first 35 ms of the IRs. As expected, the NMSE is higher than in Figure 10, since the prominent reflections constitutes a smaller fraction than in the very early part of the IRs. Despite this, the curve is of similar shape as for the 12.5 ms case, indicating that the results of the proposed method also generalizes to longer IRs.

V. Conclusion

In this work, we have defined an optimal transport distance between impulse responses and an image source model. The definition allows for a weaker formulation of the image source model that is convex and is robust to calibration errors. Using the introduced formulation, we propose a novel approach to

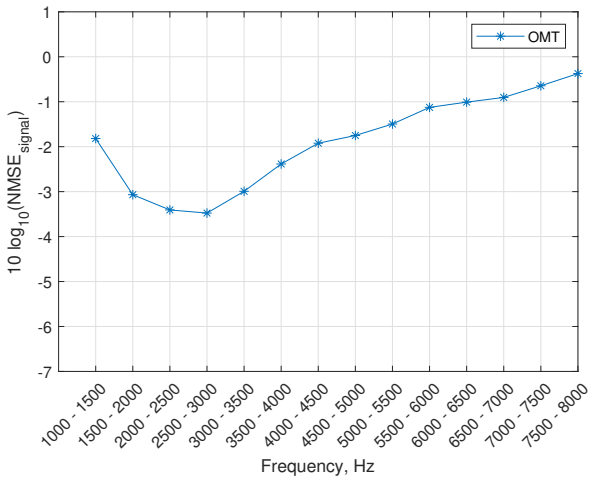


Fig. 11. The figure shows the estimated NMSE of band limited input signals using 13 sensors for IRs of length 35 ms.

interpolate the early part of impulse responses, which is robust to sensor position errors, incorporates fractional delays, allows for a linear frequency response in each reflection and the source, and incorporating sparse priors in the formulation. The proposed method is evaluated on the MeshRIR data set, clearly illustrating the method's preferable performance as compared to current state of the art room impulse response interpolation methods.

Appendix A
Upper bound of regularization parameter

An upper bound of the cost of a set of delays, as defined in (6), is here derived under the error model of calibration errors described in (12). Initially, note that the left hand side of the constraints on the sensor positions in (12) can, for any source position $\mathbf{y} \in \mathbb{R}^3$, equivalently be expressed as

$$||\tilde{\mathbf{x}}_s - \mathbf{y}||_2^2 - 2(\tilde{\mathbf{x}}_s - \mathbf{y})^T(\mathbf{x}_s - \mathbf{y}) + ||\mathbf{x}_s - \mathbf{y}||_2^2 \leq \varepsilon_x^2, \quad (23)$$

following from that

$$\begin{aligned} ||\tilde{\mathbf{x}}_s - \mathbf{x}_s||_2^2 &= ||(\tilde{\mathbf{x}}_s - \mathbf{y}) - (\mathbf{x}_s - \mathbf{y})||_2^2 \\ &= ||\tilde{\mathbf{x}}_s - \mathbf{y}||_2^2 - 2(\tilde{\mathbf{x}}_s - \mathbf{y})^T(\mathbf{x}_s - \mathbf{y}) + ||\mathbf{x}_s - \mathbf{y}||_2^2. \end{aligned}$$

Since $\mathbf{x}$ is the true sensor position, and τ_s is the delay between the true sensor position and the source, by definition

$$||\mathbf{x}_s - \mathbf{y}|| = \tau_s c. \quad (24)$$

A lower bound of the second term in (23) may thus be derived by taking the absolute value and using the Cauchy-Schwartz inequality, such that

$$\begin{aligned} -2(\tilde{\mathbf{x}}_s - \mathbf{y})^T(\mathbf{x}_s - \mathbf{y}) &\geq -2|(\tilde{\mathbf{x}}_s - \mathbf{y})^T(\mathbf{x}_s - \mathbf{y})| \\ &\geq -2||\tilde{\mathbf{x}}_s - \mathbf{y}||_2||\mathbf{x}_s - \mathbf{y}||_2. \quad (25) \end{aligned}$$

Using (24) and (25), (23) may be expressed in terms of τ_s and $a_s = ||\tilde{\mathbf{x}}_s - \mathbf{y}||_2$ such that

$$a_s^2 - 2a_s\tau_s c + \tau_s^2 c^2 \leq \varepsilon_x^2, \quad (26)$$

which is a quadratic equation in a_s with the solutions $a_s \in [\tau_s c - \varepsilon_x, \tau_s c + \varepsilon_x]$. The problem $\tilde{J}_{max}$, with the relaxed constraints in (26), is thus given by

$$\begin{aligned} \tilde{J}_{max} = \max_{a_s, \tilde{\tau}_s, s=1,\ldots,S} &\sum_{s=1}^{S} |a_s - \tilde{\tau}_s c|^p \\ \text{s.t. } a_s &\in [\tau_s c - \varepsilon_x, \tau_s c + \varepsilon_x] \\ \tilde{\tau}_s &\in [\tau_s - \varepsilon_\tau, \tau_s + \varepsilon_\tau] \quad s = 1, \ldots, S \end{aligned} \quad (27)$$

and satisfy $J_{max}(\mathbf{y}) \leq \tilde{J}_{max}$ for any $\mathbf{y} \in \mathbb{R}^3$. Furthermore, $\tilde{J}_{max}$ has the closed form solution $\tilde{J}_{max} = S(\varepsilon_x + \varepsilon_\tau)^p$.

References

[1] D. Sundström, F. Elvander, and A. Jakobsson, "Impulse response interpolation using optimal transport," in *57th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, 2023.

[2] K. Müller and F. Zotter, "Auralization based on multi-perspective ambisonic room impulse responses," *Acta Acustica*, vol. 4, 2020.

[3] T. McKenzie, N. Meyer-Kahlen, R. Daugintis, L. McCormack, S. Schlecht, and V. Pulkki, "Perceptually informed interpolation and rendering of spatial room impulse responses for room transitions," in *International Congress on Acoustics*. Gyeongju, South Korea: Acoustical Society of Korea, 2022, pp. 1–11.

[4] S. Koyama and L. Daudet, "Sparse representation of a spatial sound field in a reverberant environment," *IEEE Journal on Selected Topics in Signal Processing*, vol. 13, 2019.

[5] K. Arikawa, S. Koyama, and H. Saruwatari, "Spatial active noise control method based on sound field interpolation from reference microphone signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes island, Greece, 6 2023, pp. 1–5.

[6] H. Ito, S. Koyama, N. Ueno, and H. Saruwatari, "Spatial active noise control based on kernel interpolation with directional weighting," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2020-May, Barcelona, Spain, 5 2020, pp. 8404–8408.

[7] ——, "Feedforward spatial active noise control based on kernel interpolation of sound field," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2019-May, Brighton, UK, 5 2019, pp. 511–515.

[8] J. Brunnström, S. Koyama, and M. Moonen, "Variable span trade-off filter for sound zone control with kernel interpolation weighting," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, Singapore, 5 2022, pp. 1071–1075.

[9] T. Lee, J. K. Nielsen, J. R. Jensen, and M. G. Christensen, "A unified approach to generating sound zones using variable span linear filters," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2018-April, Calgary, Canada, 4 2018, pp. 491–495.

[10] J. Virieux, S. Operto, H. Ben-Hadj-Ali, R. Brossier, V. Etienne, F. Sourbier, L. Giraud, and A. Haidar, "Seismic wave modeling for seismic imaging," *The Leading Edge*, vol. 28, no. 5, pp. 538–544, 05 2009.

[11] A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek, and K. P. Papathanassiou, "A tutorial on synthetic aperture radar," *IEEE Geoscience and Remote Sensing Magazine*, vol. 1, no. 1, pp. 6–43, 2013.

[12] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *The Journal of the Acoustical Society of America*, vol. 138, pp. 708–730, 8 2015.

[13] A. Prodeus, M. DIdkovska, D. Motorniuk, and O. Dvornyk, "The effects of noise, early and late reflections on speech intelligibility," in *IEEE 40th International Conference on Electronics and Nanotechnology, ELNANO 2020*, Kyiv, Ukraine, 4 2020, pp. 488–492.

[14] H. Steffens, S. van de Par, and S. D. Ewert, "The role of early and late reflections on perception of source orientation," *The Journal of the Acoustical Society of America*, vol. 149, pp. 2255–2269, 4 2021.

[15] P. Samarasinghe, T. Abhayapala, M. Poletti, and T. Betlehem, "An efficient parameterization of the room transfer function," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, pp. 2217–2227, 12 2015.

[16] H.-M. Jian, Y.-S. Chen, and M. R. Bai, "Acoustic modal analysis of room responses from the perspective of state-space balanced realization with application to field interpolation," *The Journal of the Acoustical Society of America*, vol. 152, 7 2022.

[17] N. Antonello, E. D. Sena, M. Moonen, P. A. Naylor, and T. van Waterschoot, "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, pp. 1929–1941, 10 2017.

[18] J. Ribeiro, N. Ueno, S. Koyama, and H. Saruwatari, "Region-to-region kernel interpolation of acoustic transfer functions constrained by physical properties," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 30, pp. 2944–2954, 8 2022.

[19] F. Lluís, P. Martínez-Nuevo, M. B. Møller, and S. E. Shepstone, "Sound field reconstruction in rooms: Inpainting meets super-resolution," *The Journal of the Acoustical Society of America*, vol. 148, 2020.

[20] M. Hahmann and E. Fernandez-Grande, "A convolutional plane wave model for sound field reconstruction," *The Journal of the Acoustical Society of America*, vol. 152, pp. 3059–3068, 11 2022.

[21] 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, pp. 2542 – 2556, 6 2007.

[22] 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, pp. 2100–2111, 2005.

[23] R. Mignot, G. Chardon, and L. Daudet, "Low frequency interpolation of room impulse responses using compressed sensing," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, pp. 205–216, 1 2014.

[24] O. Das, P. Calamia, and S. V. Gari, "Room impulse response interpolation from a sparse set of measurements using a modal architecture," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2021-June, Toronto, Canada, 6 2021, pp. 960–964.

[25] X. Karakonstantis and E. Fernandez Grande, "Room impulse response reconstruction using physics-constrained neural networks," in *Proceedings of 10th Convention of the European Acoustics Association*. European Acoustics Association, 2023.

[26] E. Fernandez-Grande, D. Caviedes-Nozal, M. Hahmann, X. Karakonstantis, and S. A. Verburg, "Reconstruction of room impulse responses over extended domains for navigable sound field reproduction," in *Immersive and 3D Audio: from Architecture to Automotive*, Bologna, Italy, 2021, pp. 1–8.

[27] M. Pezzoli, D. Perini, A. Bernardini, F. Borra, F. Antonacci, and A. Sarti, "Deep prior approach for room impulse response reconstruction," *Sensors*, vol. 22, no. 7, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/7/2710

[28] M. Pezzoli, F. Antonacci, and A. Sarti, "Implicit neural representation with physics-informed neural networks for the reconstruction of the early part of room impulse responses," *ArXiv*, vol. abs/2306.11509, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259202816

[29] S. Dilungana, A. Deleforge, C. Foy, and S. Faisan, "Geometry-informed estimation of surface absorption profiles from room impulse responses," in *30th European Signal Processing Conference*, Belgrade, Serbia, 2022, pp. 867–871.

[30] 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 *14th International Workshop on Acoustic Signal Enhancement, IWAENC*, Antibes, France, 9 2014, pp. 114–118.

[31] ——, "Evaluation of a numerical method for identifying surface acoustic impedances in a reverberant room," in *Euronoise*, Maastricht, Netherlands, 5 2015.

[32] 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, 3 2009.

[33] N. Bertin, S. Kitić, and R. Gribonval, "Joint estimation of sound source location and boundary impedance with physics-driven cosparse regularization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, 05 2016, pp. 6340–6344.

[34] W. Jin and W. B. Kleijn, "Theory and design of multizone soundfield reproduction using sparse methods," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, pp. 2343 – 2355, 12 2015.

[35] S. Koyama and L. Daudet, "Comparison of reverberation models for sparse sound field decomposition," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, vol. 2017-October, New Paltz, NY, USA, 10 2017, pp. 214–218.

[36] A. Moiola, R. Hiptmair, and I. Perugia, "Vekua theory for the helmholtz operator," *Zeitschrift fur Angewandte Mathematik und Physik*, vol. 62, 2011.

[37] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[38] I. Tsunokuni, K. Kurokawa, H. Matsuhashi, Y. Ikeda, and N. Osaka, "Spatial extrapolation of early room impulse responses in local area using sparse equivalent sources and image source method," *Applied Acoustics*, vol. 179, 2021.

[39] G. H. Koopmann, L. Song, and J. Fahnline, "A method for computing acoustic fields based on the principle of wave superposition," *Journal of the Acoustical Society of America*, vol. 86, pp. 2433–2438, 1989.

[40] M. E. Johnson, S. J. Elliott, K.-H. Baek, and J. Garcia-Bonito, "An equivalent source technique for calculating the sound field inside an enclosure containing scattering objects," *The Journal of the Acoustical Society of America*, vol. 104, pp. 1221–1231, 1998.

[41] I. Dokmanic, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, pp. 12 186–12 191, 7 2013.

[42] L. Remaggi, P. J. Jackson, P. Coleman, and W. Wang, "Acoustic reflector localization: Novel image source reversion and direct localization methods," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, 2017.

[43] M. Crocco, A. Trucco, and A. Del Bue, "Uncalibrated 3d room geometry estimation from sound impulse responses," *Journal of the Franklin Institute*, vol. 354, no. 18, pp. 8678–8709, 2017.

[44] L. Remaggi, P. Jackson, P. Coleman, and T. Parnell, "Estimation of object-based reverberation using an ad-hoc microphone arrangement for live performance," in *Audio Engineering Society Convention 144*. Audio Engineering Society, 05 2018.

[45] D. Hu, Z. Chen, and F. Yin, "Passive geometry calibration for microphone arrays based on distributed damped newton optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 118–131, 2021.

[46] M. Leigsnering, F. Ahmad, M. Amin, and A. Zoubir, "Multipath exploitation in through-the-wall radar imaging using sparse reconstruction," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 50, pp. 920–939, 2014.

[47] T. Sprunck, A. Deleforge, Y. Privat, and C. Foy, "Gridless 3d recovery of image sources from room impulse responses," *IEEE Signal Processing Letters*, vol. 29, pp. 2427–2431, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251929009

[48] M. Kreković, I. Dokmanić, and M. Vetterli, "Echoslam: Simultaneous localization and mapping with acoustic echoes," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2016-May, Shanghai, China, 3 2016, pp. 11–15.

[49] M. Krekovic, I. Dokmanic, and M. Vetterli, "Shapes from echoes: Uniqueness from point-to-plane distance matrices," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2480 – 2498, 3 2020.

[50] R. Scheibler, I. Dokmanic, and M. Vetterli, "Raking echoes in the time domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2015-August, South Brisbane, QLD, Australia, 8 2015, pp. 554–558.

[51] P. Blondel, "The handbook of sidescan sonar," pp. 249–276, 1 2009.

[52] M. P. Hayes and P. T. Gough, "Synthetic aperture sonar: A review of current status," *IEEE Journal of Oceanic Engineering*, vol. 34, pp. 207–224, 2009.

[53] T. Inouye, K. Shinosaki, A. Iyama, and Y. Matsumoto, "Localization of activated areas and directional eeg patterns during mental arithmetic," *Electroencephalography and Clinical Neurophysiology*, vol. 86, pp. 224–230, 1993.

[54] T. 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, pp. 2301–2312, 11 2013, exploiting sparse structure in time-domain.

[55] S. Damiano, F. Borra, A. Bernardini, F. Antonacci, and A. Sarti, "Soundfield reconstruction in reverberant rooms based on compressive sensing and image-source models of early reflections," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, vol. 2021-October, New Paltz, NY, USA, 10 2021, pp. 366–370.

[56] F. Katzberg, R. Mazur, M. Maass, M. Bohme, and A. Mertins, "Spatial interpolation of room impulse responses using compressed sensing," in *16th International Workshop on Acoustic Signal Enhancement, IWAENC 2018*, Tokyo, Japan, 9 2018, pp. 426–430.

[57] S. Siltanen, T. Lokki, S. Tervo, and L. Savioja, "Modeling incoherent reflections from rough room surfaces with image sources," *The Journal of the Acoustical Society of America*, vol. 131, pp. 4606–4614, 6 2012.

[58] Q. Feng, F. Yang, and J. Yang, "Interpolation of the early part of the acoustic transfer functions using block sparse models," *The Journal of the Acoustical Society of America*, vol. 142, 12 2017.

[59] C. Villani, *Optimal Transport: Old and New*. Springer, 2008.

[60] G. Peyré and M. Cuturi, "Computational optimal transport: With applications to data science," *Foundations and Trends in Machine Learning*, vol. 11, pp. 355–607, 2019.

[61] T. T. Georgiou, J. Karlsson, and M. S. Takyar, "Metrics for power spectra: An axiomatic approach," *IEEE Transactions on Signal Processing*, vol. 57, pp. 859 – 867, 10 2009.

[62] F. Elvander, A. Jakobsson, and J. Karlsson, "Interpolation and extrapolation of toeplitz matrices via optimal mass transport," *IEEE Transactions on Signal Processing*, vol. 66, pp. 5285 – 5298, 10 2018.

[63] F. Elvander, "Estimating inharmonic signals with optimal transport priors," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, 2023, pp. 1–5.

[64] F. Elvander and A. Jakobsson, "Defining fundamental frequency for almost harmonic signals," *IEEE Transactions on Signal Processing*, vol. 68, pp. 6453 – 6466, 11 2020.

[65] F. Elvander, I. Haasler, A. Jakobsson, and J. Karlsson, "Multi-marginal optimal transport using partial information with applications in robust localization and sensor fusion," *Signal Processing*, vol. 171, 6 2020.

[66] ——, "Tracking and sensor fusion in direction of arrival estimation using optimal mass transport," in *European Signal Processing Conference*, vol. 2018-September, Rome, Italy, 8 2018, pp. 1617–1621.

[67] F. Elvander, J. Karlsson, and T. van Waterschoot, "Convex clustering for multistatic active sensing via optimal mass transport," in *29th European Signal Processing Conference*, Dublin, Ireland, 2021, pp. 1730–1734.

[68] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde, "Optimal mass transport: Signal processing and machine-learning applications," *IEEE Signal Processing Magazine*, vol. 34, pp. 43 – 59, 7 2017.

[69] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, "Scaling algorithms for unbalanced optimal transport problems," *Mathematics of Computation*, vol. 87, pp. 2563–2609, 5 2018.

[70] B. Pass, "Multi-marginal optimal transport: Theory and applications," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 49, pp. 1771–1790, 2015.

[71] A. Geldert, N. Meyer-Kahlen, and S. J. Schlecht, "Interpolation of spatial room impulse responses using partial optimal transport," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. Rhodes, Greece: IEEE, 6 2023, pp. 1–5.

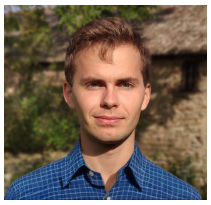
[72] M. Larsson, V. Larsson, K. Åstrom, and M. Oskarsson, "Optimal trilateration is an eigenvalue problem," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2019-May, Brighton, UK, 5 2019, pp. 5586–5590.

[73] S. Koyama, T. Nishida, K. Kimura, T. Abe, N. Ueno, and J. Brunnstrom, "Meshrir: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, vol. 2021-October, New Paltz, NY, USA, 8 2021, pp. 1–5.

[74] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, "Splitting the unit: Delay tools for fractional delay filter design," *IEEE Signal Processing Magazine*, vol. 13, pp. 30 – 60, 1 1996.

[75] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, ver 2.1," 2017.

[76] S. G. McGovern, "Fast image method for impulse response calculations of box-shaped rooms," *Applied Acoustics*, vol. 70, no. 1, pp. 182–189, 2009.



David Sundström (Student member, IEEE) was born in Stockholm, Sweden, 1999. He received his M.Sc. degree in engineering physics from Lund University (LU), Sweden, in 2022, and is currently working towards a Ph.D. in Mathematical Statistics at LU. In spring 2024, he was a visiting researcher at NII S. Koyama's Lab, Tokyo, Japan. His research interests include statistical signal processing, acoustic inverse problems, and related applications in spatial audio signal processing.

Filip Elvander (Member, IEEE) received the M.Sc. degree in industrial engineering and management and the Ph.D. degree in mathematical statistics from Lund University, Lund, Sweden, in 2015 and 2020, respectively. He has been a Postdoctoral Research Fellow with the Stadius Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium, and with the Research Foundation – Flanders (FWO). He is currently an Assistant Professor of signal processing with the Department of Information and Communications Engineering, Aalto University, Finland. He is a Member of the EURASIP Technical Area Committee on Signal and Data Analytics for Machine Learning. His research interests include inverse problems, robust estimation, and convex modeling and approximation techniques in statistical signal processing and spectral analysis.

Andreas Jakobsson (Senior Member, IEEE) received his M.Sc. from Lund Institute of Technology and his Ph.D. in Signal Processing from Uppsala University in 1993 and 2000, respectively. Since, he has held positions with Global IP Sound AB, the Swedish Royal Institute of Technology, King's College London, and Karlstad University, as well as held an Honorary Research Fellowship at Cardiff University and a guest professorship at Harbin Engineering University. He is the co-founder of four start-up companies, has been a visiting researcher at King's College London, Brigham Young University, Stanford University, KU Leuven, and University of California, San Diego, as well as acted as an expert for the IAEA. He is currently Professor of Mathematical Statistics at Lund University, Sweden. He has published his research findings in more than 300 refereed journal and conference papers, and has filed six patents. He has also authored a book on Time Series Analysis (Studentlitteratur), and co-authored (together with M. G. Christensen) a book on Multi-pitch Estimation (Morgan & Claypool). He is a member of the Royal Swedish Physiographic Society, a Senior Member of IEEE, as well as a Subject Editor for Elsevier Signal Processing. He has previously also been a Senior Associate Editor for IEEE Transactions on Signal Processing (2018-2022), a member of the EURASIP Special Area Team on Signal Processing for Multisensor Systems (2015-2021), a member of the IEEE Sensor Array and Multichannel (SAM) Signal Processing Technical Committee (2008-2013), an Associate Editor for the IEEE Transactions on Signal Processing (2006-2010), the IEEE Signal Processing Letters (2007-2011), the Research Letters in Signal Processing (2007-2009), and the Journal of Electrical and Computer Engineering (2009-2014). His research interests include statistical and array signal processing, detection and estimation theory, and related applications in remote sensing, telecommunication, and biomedicine.