Compressed sensing of impulse responses in rooms of unknown properties and contents

Elias Zea

KTH Royal Institute of Technology, The Marcus Wallenberg Laboratory for Sound and Vibration Research, Teknikringen 8, SE-100 44 Stockholm, Sweden

Abstract

This paper introduces a method to recover unmeasured room impulse responses (RIRs) in acoustical spaces with unknown properties and contents, by means of a compressed sensing methodology. Methods published in the existing literature have been validated in empty, convex rooms; a limited subset of the many, diverse acoustical spaces one can encounter. It results a challenge to represent such diverse wave phenomena with a sparse set of plane waves or equivalent sources, given the coupling between the sparsity of such representations and hypotheses regarding the properties of the acoustical space and its contents, far-field measurement distances, and other parameters. In contrast to this philosophy, the method introduced in this paper exploits the sparsity inherent to the mathematical structure of the wavefronts present in the RIRs, which without further hypotheses carry themselves all the information about the wave propagation in the room. In essence, the measured RIRs are instead represented with a sparse set of curved elementary functions of various sizes, propagation directions and times of arrival, which are linked with the various shapes and locations of the unknown scatterers and boundaries in the room. The main contribution of this work is thus to enable the measurement of RIRs in more complex acoustical spaces, while keeping the number of microphones to a minimum with the use of compressed sensing. The method is formulated as a sparse optimization problem, and the solution is obtained with an iterative thresholding algorithm whose threshold value is determined from the measurements. An analysis of sensing coherence is included, and the performance of the method is experimentally evaluated with 1D microphone array measurements in two lecture rooms and one meeting room. For the sake of comparison, the RIRs are also linearly interpolated using a low-pass filter in the wavenumber-frequency domain. The experimental results demonstrate that the proposed method is superior than linear interpolation in all the cases investigated, motivating further development of the method to higher spatial dimensions. In terms of accuracy, the proposed method attains recovery errors in the same order of magnitude as those attained by methods in the literature, yet here the acoustical spaces have arbitrary contents and exhibit more complex geometries and boundary conditions.

Keywords: room impulse responses, compressed sensing, interpolation, complex acoustical spaces, shearlet dictionaries

1. Introduction

Room impulse response (RIR) measurements comprise a conventional routine in the acoustical analysis of cabins, rooms, halls, auditoria, and the like. RIRs can be thought of as acoustical imprints of a given space, which on the one hand allows for the quantification of representative indicators [1–3] for design considerations or acoustic treatment, such as sound compensation, control, or enhancement; and on the other hand serves as input data in –to mention a few– auralization, teleconferencing, and augmented/virtual reality applications.
RIR measurements involve recording the pressure response of the room to an impulsive sound source, at a number of positions with microphones. Faced with the challenge of acquiring meaningful information with limited resources, the acoustician has to compromise between the spatial resolution and the effort it takes to record the data.

Generally speaking, it is impractical to measure the responses in the whole room within the entire human audible range. When the source and the microphones are placed apart by a distance significantly larger than the acoustic wavelength $\lambda$, the distance (or spatial resolution) between microphones $d$ is governed by the Nyquist-Shannon (NS) sampling theorem: $d < \lambda/2$ (as little as 0.85 cm for frequencies up to 20 kHz). When the source and the microphones are not many wavelengths apart, the sound field is further composed of high-resolution near-field waves, and $d$ ought to be smaller to comply with the NS theorem and ensure a sufficiently large signal-to-noise ratio [5]. In this range of small wavelengths, the acoustic waves also interact with the microphones, adding scattering distortions to the measured sound field. Nevertheless, even if none of these aspects would be of concern, the number of microphone positions ultimately depends on how large a measurement space the acoustician wishes to sample: for frequencies within the human audible range, this would theoretically exceed 13 thousand positions per square meter.

To avoid carrying out such a vast number of measurements, acousticians have used computers since the 60’s to approximate sound fields in rooms in a variety of ways [6], relying on the knowledge of the room properties as well as its contents. A central result is the Schroeder frequency [7], establishing two regimes for wave behavior according to the modal density of the room. As with light waves, acoustic waves above this frequency can be treated geometrically. Research in this direction has resulted in the development of geometrical acoustics methods [8], such as the image source model [9] and ray-tracing [10]. Besides compromising their validity below Schroeder’s limit, classical geometrical methods are inaccurate in the presence of diffraction and scattering phenomena [8], caused by objects present in the room for instance. In such cases, closer approximations may result with integral diffraction formulations [11], wave-based models [12, 13], and hybrid methods [14, 15]. As anticipated earlier, however, accurate computer predictions rely on accurate modelling —that is, accurate knowledge— of the room properties and contents [16].

After more than five decades of computer predictions [17], measurement-based methods are experiencing a revival with the introduction of compressed sensing (CS) [18, 19], a framework to recover signals sampled below NS rates. The success of CS relies on signal sparsity in a transform space (dictionary), and on a highly incoherent sensing operation. The benefit this framework has brought to the acoustics community is a reduction of microphones in various data acquisition and signal processing tasks [20], one of them being the task of interpolating RIRs. One of the pioneering papers was written by Mignot et al. [21], in which a dictionary of virtual monopole sources is used to interpolate the early part of the RIRs —thereby exploiting the temporal sparsity of the early part, assuming the room is empty and there is no diffraction phenomena. A few years later, Antonello et al. investigated the use of a dictionary of time-dependent equivalent sources in various sparsity-promoting domains [22], assuming far-field measurement distances, and performed an experimental validation in an empty, rectangular room. In that paper, the authors show that adding prior information about the sound source in the model improves the accuracy of the interpolation [22]. Modal expansions [23] and plane-wave approximations [24–26] have also been used to interpolate and extrapolate the sound field, being accurate in convex rooms and at frequencies below Schroeder’s limit: the sparsity of waves is linked with a low modal density. Another recent approach proposed by Katzberg et al. interpolates the sound field with a moving microphone, validated numerically in an empty rectangular room [27]. Amidst the valuable efforts spent in this direction, there are to the author’s knowledge no published papers in the literature that have investigated the task of interpolating RIRs in rooms with unknown properties and contents.

In this way, it is the main goal of this paper to introduce and validate a method for interpolating RIRs, measured in rooms with unknown properties and contents, thereby keeping the number of microphones to a minimum. In contrast to signal representations whose sparsity is coupled with hypotheses regarding the measurement conditions (e.g. convexity of the room or its contents), the method presented here makes use of multi-scale directional representations of the RIRs, whose sparsity is coupled with the mathematical structure of the wavefronts present in the data. Examples of multi-scale directional dictionaries are: curvelets [28], shearlets [29], and directional filter banks [30]. These dictionaries, commonly used by image processors,
are redundant and consist of curved elementary functions of various scales, orientations, and translations, which—in the context of RIR measurements—can be understood as wavefronts of various sizes, propagation directions, and times of arrival. In particular, the method introduced here uses 2D shearlet dictionaries to interpolate RIRs measured with linear (1D) microphone arrays; however, it can be extended to planar (2D) microphone array data by using 3D shearlet dictionaries [31]. The method can to a considerable extent be regarded as a room acoustics equivalent of a former method to interpolate missing seismic traces with curvelets [32]. The interpolated solution is found via iterative thresholding of the shearlet coefficients, and the performance is experimentally evaluated with RIRs measured in three different rooms. A comparison is also drawn with linear interpolation using a low-pass filter, which is in all cases outperformed by the proposed method.

The manuscript is organized as follows. Section 2 presents theoretical background suggesting that wavefronts in rooms can be sparsely represented with shearlets, followed by the formulation of the interpolation problem as an optimization task. The sensing coherence of the problem is analyzed in Sec. 3, and the computer implementation is outlined in Sec. 4. Section 5 presents the experimental validation and discusses the results, including the comparison with the linear interpolation method, an analysis of computational aspects, and the influence of noise levels in the recovery. Conclusions are drawn in Sec. 6.

2. Theory

Spatio-temporal RIR data is composed of a free-field wave that experiences multiple acoustic distortions due to the presence of walls, furniture, windows, etc. This can be observed in the RIR image in Fig. 1, in which a group of wavefronts gradually blend and vanish in time as the room scatters and dissipates the acoustic energy. These RIR images consist of the sound pressure plotted against time and space variables, which results from recordings of $T$ time samples with microphones at $M$ locations in space.

![Figure 1: RIR image of size $m = 350 \times 100$ in a meeting room.](image)

2.1. Sparse wavefront representations

The central idea is to represent (decompose) the RIR image into a sparse collection of curved elementary functions, known as shearlets [33], which closely resemble wavefronts of various sizes, propagation directions and times of arrival. One can think of shearlets as some form of directional wavelets. In the context of image processing, shearlets (as well as curvelets) provide the sparsest representation of a discrete image signal $\mathbf{y} = \{y_1, \ldots, y_m\}$ that is composed of $C^2$ functions smooth away from $C^2$ discontinuities (e.g. curved
edges in an image). In fact, the error of approximating such a signal \( y \) with another signal \( y^* = \{ y_1^*, \ldots, y_m^* \} \) that is synthesized from the \( L \) largest shearlet dictionary coefficients follows [34]

\[
\sum_i |y_i - y_i^*|^2 = \mathcal{O} \left( L^{-2} (\log L)^3 \right), \quad \text{as } L \to \infty.
\]

(1)

To this day this is the closest to the optimal rate \( \mathcal{O} (L^{-2}) \) for this particular class of functions [35]. This essentially means that an optimally sparse approximation of any such function in the signal \( y \) can be attained by using shearlets. This approximation is superior compared with that of Fourier bases \( \mathcal{O} (L^{-1/2}) \) and wavelets \( \mathcal{O} (L^{-1}) \), as their multivariate expansions are not generated from anisotropic elements [35].

To put these results in the context of RIR images, it is valuable to develop some intuition regarding the presence of such classes of functions in RIR data. One argument is the resemblance of the spatio-temporal wavefronts, such as those in Fig. 1, to curved discontinuities of different sizes and orientations, and the resemblance of the spaces between wavefronts to smooth functions. Another argument is based on the response of the linearized wave equation to an impulse (Dirac) source term \( \delta(x, t) \), in a medium with sound speed \( c \) and unknown boundary conditions,

\[
\frac{\partial^2 \varphi}{\partial t^2} - c^2 \nabla^2 \varphi = \delta(x, t),
\]

(2)

whose solutions \( \varphi(x, t) \) should be (at least) \( \mathcal{C}^2 \) in order for the second-order time \( \partial^2 / \partial t^2 \) and space \( \nabla^2 \) derivatives to exist. In addition, these ideas have been inspired on previous papers on the recovery of missing seismic traces with curvelets [32] and shearlets [36, 37]; the literature of seismic models applied in room acoustics [12, 38, 39]; and the remarkable work by Candès and Demanet proving the optimality of curvelets to sparsely represent linear operators in hyperbolic differential equations (such as the Green’s function in the wave equation) [40]; all of which strongly suggests that these curved elements can provide a sparse approximation of acoustic wavefronts in rooms.

In this paper, the chosen dictionary is composed of 2D cone-adapted shearlets [33], making use of a Matlab implementation of the fast finite shearlet transform [41, 42]. With some careful considerations, the exposed mathematics are compatible with other multi-scale directional dictionaries, as long as these are equipped with explicit forward and inverse transform operations. In particular, the method presented here makes use of shearlets as sparse representations of RIRs that are measured with linear (1D) microphone arrays.

2.2. Shearlet decompositions of RIR images

Let us begin by arranging the 2D RIR image in the column vector \( y \in \mathbb{R}^m \), with \( m = T \times M \) as the total amount of time samples and microphone positions. Then, the discrete shearlet coefficients \( \alpha \in \mathbb{C}^n \) are obtained via

\[
\alpha = S y,
\]

(3)

where the block matrix

\[
S = \begin{bmatrix}
S_1 \\
\vdots \\
S_K
\end{bmatrix} \in \mathbb{C}^{n \times m}
\]

(4)

encompasses the 2D discrete shearlet transform operation. The shearlet transform is in fact a redundant transform, as it decomposes the image into

\[
K = 1 + \sum_{\sigma=1}^{\tau} 2^{\sigma+1}
\]

(5)

representations [41], i.e. \( n = K \times m \) shearlet coefficients. A larger number of scales \( \tau \) is generally linked with a broader variety of shapes that can be represented. Examples of shearlet elements are shown in Fig. 2, drawn from a dictionary of \( \tau = 3 \) scales.
Figure 2: Space-time shearlets drawn from a dictionary of \( \tau = 3 \) decomposition scales [41]. With respect to Eq. (4), the shearlets are indexed with (a) \( k = 1 \), (b) \( k = 5 \), (c) \( k = 9 \), and (d) \( k = 14 \), corresponding to scales (a) \( \sigma = 0 \), (b) \( \sigma = 1 \), (c) \( \sigma = 2 \), and (d) \( \sigma = 3 \). In this particular example: \( T = 128 \), \( M = 128 \), the sampling frequency \( f_s = 11250 \) Hz, and the microphone spacing \( d = 3 \) cm. Figure (c) shows the size of the shearlet, which follows the anisotropic scaling law: width = length \(^2 \) [33].

It shall be stressed at this point that, although the computational complexity increases, using redundant representation systems brings benefits that have been well studied in the literature [43]: increased robustness of the sensing mechanism against noise and missing data, as well as a more flexible representation of the signals. Also, it will later be shown in Sec. 5.2 that the computation times are quite reasonable.

In Eq. (4), every block \( S_k \in \mathbb{C}^{m \times m} \) has Toeplitz structure due to translation-invariance, thus the spatio-temporal convolutions of Eq. (3) can be performed in Fourier space. In fact, the operation carried out by \( S \) is in practice done with fast Fourier transforms (FFTs) and multiplications between the signal and shearlet spectra [41, 42]. The shearlet spectrum associated with the \( k \)-th block of \( S \), as depicted for various indices \( k \) in Fig. 3, is composed of a group of plane waves whose wavenumber-frequency (\( \psi, f \)) content depends on the shearlet scale \( \sigma \), and on the shearlet orientations within that particular scale. For a precise definition of these spectra the reader is referred to the paper by Haüser and Steidl [41]. There is a total of \( 2^{\sigma+1} \) shearlet orientations at the \( \sigma \)-th scale, which together span a corona \( W_\sigma \), with \( 1 \leq \sigma \leq \tau \). An example of such a corona is shown in Fig. 3(c) for the scale \( \sigma = 2 \). The inner and outer sides of the corona follow from a dyadic partitioning of the Fourier domain [33], and can be linked with the sampling parameters of the image [41]. The coarsest shearlet \( \sigma = 0 \), depicted in Fig. 3(a), is a low-pass filter function with no orientation, and its wavenumber-frequency spectrum is flat within a square region.

From a physical viewpoint, the shearlet decomposition of a RIR image via Eq. (3) involves an approximation of the width of the wavefronts with different shearlet scales; the propagation directions of the wavefronts with different shearlet orientations; and the spatio-temporal locations of the wavefronts with different shearlet translations. This is equivalent to considering multiple spatio-temporal domains, as shown in Fig. 4 below, each being a \( T \times M \) representation of the RIR image with a particular kind of information.

The cone-adapted shearlets constitute a Parseval frame [43], so the adjoint (dictionary) matrix

\[
D = \begin{bmatrix} D_1 & \cdots & D_K \end{bmatrix} = S'^H
\]

synthesizes the signal back from its coefficients

\[
y = D \alpha.
\]
Figure 3: Magnitude of the wavenumber-frequency \((\psi, f)\) spectra of shearlets in correspondence with Fig. 2. In (c), the corona \(W_2\) contains \(2^{2+1} = 8\) shearlet orientations.

Figure 4: Decomposition of RIR image from Fig. 1 into shearlets \(k = 12\) and \(26\), with similar orientation but different scales \(\sigma = 2\) and \(3\).

2.3. Interpolation problem

As the interest lies in recovering microphone responses that have not been measured, the problem can be posed as an interpolation task. Let us begin by under-sampling the RIR image in a non-uniform (pseudorandom) fashion, by an integer factor, say \(u = m/q\), as

\[
\hat{y} = \Lambda y,
\]

where the mask \(\Lambda \in \mathbb{R}^{q \times m}\), also known as selection operator, is an identity matrix with \(r = m - q\) rows removed corresponding to the \(T\) time samples at \(R\) missing microphone positions. Visually this means the
degradation of Fig. 1 into Fig. 5, and mathematically that $r = T \times R$ samples are missing.

Figure 5: Masking $R = 67$ out of $M = 100$ microphone recordings of the RIR image in Fig. 1.

Substituting Eq. (3) into (8) results in the linear under-determined system of equations

$$\hat{y} = \Phi \alpha + e,$$

where the sensing matrix $\Phi = \Lambda D \in \mathbb{C}^{q \times n}$, and the column vector $e$ contains errors due to model misfit and measurement noise.

This noisy inverse problem can be formulated as an optimization task, by means of penalizing the misfit with the model and the number of shearlets used to recover $y$. The solution can be found by optimizing [44]

$$(P_0) : \min_{\alpha} \|\alpha\|_0 \quad \text{subject to} \quad \|\hat{y} - \Phi \alpha\|_2 \leq \|e\|_2$$

(10)

where $\|\alpha\|_0$ is the number of non-zero shearlets, and the $\ell_p$-norm $\| \cdot \|_p = (\sum |\cdot|^p)^{1/p}$ for $1 \leq p < \infty$. Solving (P0) directly entails a combinatorial search that can be solved with matching pursuit algorithms [44]; however it can soon become intractable because of the (typically large) dimensions of the problem—all the more if the columns of $\Phi$ cannot be defined with FFTs. In that case it is common to relax and unconstrain the optimization functional, by means of penalizing the $\ell_1$-norm with an appropriate Lagrange multiplier $\zeta$ [44]

$$(P_1) : \min_{\alpha} \frac{1}{2} \|\hat{y} - \Phi \alpha\|_2^2 + \zeta \|\alpha\|_1.$$ (11)

Efficient ways to solve (P1) are, for example, interior-point methods [45] or iterative thresholding [46]. But in order to be close enough to the true solution, it is crucial to promote sensing incoherence.

3. Coherence analysis

To recover the missing information via (P1) with high probability [19], the sensing operation carried out by $\Phi$ ought to be highly incoherent—in practice linked with the randomness of the sampling scheme. For example, regular under-sampling by a factor $u = M/(M - R)$ keeps a microphone every $\lfloor u \rfloor$ positions, which is known to introduce aliasing artifacts in the wavenumber-frequency spectrum [47]. Instead, random under-sampling dissipates the spatial aliasing into seemingly uncorrelated noise [32, 48, 49], but the gap between samples cannot be arbitrarily large.
3.1. Gap size vs. randomness

In theory, recoveries are accurate when the gap size is (at most) asymptotically smaller \([37, 50]\) than the length of the shearlet elements. This can be understood geometrically by looking at Fig. 6: an insufficiently long element may fail to interpolate the gap. As illustrated in Fig. 2(c), the length of a cone-adapted shearlet element follows the anisotropic scaling rule \(\text{width} \propto \text{length}^2\) \([33]\); and it varies with the discretization of the dictionary: sampling frequency \(f_s\), microphone spacing \(d\), and number of scales \(\tau\). Manipulating these variables, however, can lead to practical complications (e.g. arrays that are too large), and one should look for a compromise.

![Figure 6: Shearlet interpolating three different gap sizes. In this particular sketch, the shearlet element (oval) is likely to fail at interpolating the widest gap due to insufficient length.](image)

In this work we shall adopt an under-sampling strategy suggested by Hennenfent and Herrmann, called jittered (pseudorandom) under-sampling \([48]\), which aims at regulating the size of the maximum gap. The microphone positions are first regularly under-sampled, and thereafter perturbed with integers \(g_1, \ldots, g_{M-P}\), \(g_i \in [-u, u]\), drawn from a random uniform discrete distribution. It should be mentioned that the accuracy of solving \((P_1)\) can be upper-bounded in the continuous regime \([36, 37]\). Nevertheless, today’s lack of mathematical tools precludes a derivation of such bounds in the discrete regime, and one has to take other measures to assess the sensing properties of the system.

3.2. Quantifying coherence

The sensing coherence of the system can generally be quantified with the linear dependency between distinct columns of \(\Phi \in \mathbb{C}^{q \times n}\),

\[
\mu(\Phi) = \max_j \max_i |\langle \phi_i, \phi_j \rangle|,
\]

provided \(\langle \cdot, \cdot \rangle\) is the inner product and the columns \(\phi_1, \ldots, \phi_n\) are normalized.

Given the typical dimensions of RIR images (from thousands to millions of samples), and the redundancy of the shearlet transform \(K\) (e.g. 29 for \(\tau = 3\)), the computation of \(\mu\) can easily become intractable. It is however possible to exploit the geometrical features of the mask \(\Lambda\), to the end of quantifying the coherence of \(\Phi\) using a smaller set of columns. To explain this, consider Fig. 5. The vertical stripes aligned with the time axis (caused by the masked microphone positions) are most correlated with vertically oriented shearlets [e.g. Fig. 2(d)]; whereas the actual sound pressure wavefronts are most correlated with horizontally oriented shearlets [e.g. Fig. 2(c)]. In this way, it is possible to determine which shearlets represent artificial (and which represent natural) parts of the RIR image, based on the degree of coherence the shearlets have with respect to the vertical stripes. The computation of the cluster coherence is supported by this logic \([51]\),

\[
\mu_c(\Gamma, \Phi) = \max_j \sum_{i \in \Gamma} |\langle \phi_i, \phi_j \rangle| = \max_j \vartheta_\Gamma(j),
\]

which emphasizes the coherence of the columns of \(\Phi\), belonging to the cluster \(\Gamma \subset \{1, \ldots, n\}\), that interact the most with the mask \(\Lambda\). As suggested in the seismic interpolation problem \([32]\), the most coherent columns
of $\Phi$ should be removed from the sensing operation, as they only introduce artifacts in the recovered images. How these columns are chosen in this paper is the purpose of the following section. The approach is to compute $\vartheta_\Gamma$ with a cluster $\Gamma$ defined by the columns of $\Phi$ corresponding to strictly vertical shearlets [such as that in Fig. 2(d)], and restrict the columns of $\Phi$ that are most coherent with said strictly vertical shearlets.

3.3. Forbidden shearlets

We shall start by defining the columns of $\Phi$ in terms of shearlet elements $k = 1, \ldots, K$, their scales $\sigma = 0, 1, \ldots, \tau$, and translations $1, \ldots, m$. Provided a cone-adapted shearlet dictionary of $\tau$ scales, the $m$ translations of the $k$-th shearlet element are allocated in the columns $j_k \in \mathbb{N}^m$

$$j_k = \{(k - 1)m + 1, \ldots, km\}.$$  

That is, for instance, $j_1$ corresponds to the coarsest shearlet $k = 1$, whose translations are allocated in the columns $1, \ldots, m$ of $\Phi$. In accordance with the arrangement of indices $k$ in the fast finite shearlet transform (see Fig. 11 in [41]), the columns of $\Phi$ corresponding to the strictly vertical shearlets can be defined for all translations as

$$\Gamma_V = \bigcup_{\sigma=1}^{\tau} j_k, \quad k = 2^{\sigma+1} - 2.$$  

Then, computing the values $\vartheta_{\Gamma_V}(j_k)$, obtained by setting $\Gamma = \Gamma_V$ in Eq. (13), points out the vertical shearlets [other than the strictly vertical ones in Eq. (15)] that should be removed. In other words, higher coherence values of shearlets with respect to the cluster of strictly vertical shearlets $\Gamma_V$, are associated with higher chances of introducing artificial information in the recovered images.

Figure 7 shows an example of computing $\vartheta_{\Gamma_V}(j_k)$ given $\tau = 3$ scales and $m = 1024$ translations. As it can be seen, the values of $\vartheta_{\Gamma_V}(j_k)$ have some form of coherence floor on every scale (enclosed with dashed boxes). On this floor lie the least coherent shearlets, here denoted with $\Gamma_{LC} \in \mathbb{N}^{m(1+2\tau)}$, which carry natural information for being aligned with the acoustic wavefronts. In this way, the set of forbidden shearlets—carrying artificial information—can be defined as

$$\Gamma_{MC}(\sigma) = \Gamma_{\sigma} \setminus \Gamma_{LC}(\sigma),$$  

![Figure 7: Coherence values $\vartheta_\Gamma(j_k)$ for a cluster $\Gamma = \Gamma_V$, at shearlet scales (a) $\sigma = 0$, (b) $\sigma = 1$, (c) $\sigma = 2$, and (d) $\sigma = 3$, provided $\tau = 3$ and $m = 1024$. In (c), the sets $\Gamma_{LC} = j_6 \cup \ldots \cup j_{12}$ and $\Gamma_{MC} = j_6 \cup j_{7} \cup j_{13}$. For the sake of easier presentation, the abscissa is parametrically defined with the index $k$, and for each $k$ there are $m$ translations as defined in Eq. (14).](image-url)
where $\Gamma_\sigma \in \mathbb{N}^{m(2\sigma+1)}$ denotes the set of all columns for any given $\sigma \geq 1$. This is to say, $\Gamma_{MC}$ contains the $m(2\sigma - 1)$ columns complementary to $\Gamma_{LC}$ at the $\sigma$-th scale [e.g. see the enclosing ovals in Fig. 7(c)].

Besides this mathematical analysis, there is in fact a physical interpretation of this restriction in terms of the plenacoustic function (PAF) [47], which is a multivariate pressure function of time and space in a given acoustic environment. As illustrated with the hourglass in Fig. 8, the wavenumber-frequency spectrum of the 2D PAF in the far-field is band-limited [47] in the symmetric cone $\mathcal{P} = \{(\psi, f) \in \mathbb{R}^2 : |\psi| \leq 2\pi f/c\}$. In this way, the PAF is almost complementary in spectrum to the most coherent shearlets $\Gamma_{MC}$ (numbered wedges). It is almost complementary because the dictionary partition (dash-dotted lines) almost never coincides with the far-field PAF (hourglass), which leaves the space $\Gamma_\sigma \cap (\Gamma_{MC} \cup \mathcal{P})$ for near-fields to be represented as well (striped regions).

![Figure 8: Wavenumber-frequency (ψ, f) spectra of the far-field plenacoustic function P (hourglass), the restricted shearlets Γ_{MC} (numbered wedges), a 2-scale cone-adapted shearlet partition [41] (dash-dotted lines), and the near-field partition (striped regions).](image)

The restriction of the shearlets is carried out with an expansion matrix $E$, which, similarly but opposite to the selection matrix $A$, is an identity matrix with columns removed according to the indices $k$ satisfying $j_k \subset \Gamma_{MC}$. This is done by rewriting the sensing matrix as $\Phi = \Lambda \tilde{D} = \Lambda DE$. In other words, $\alpha$ has now $K \times m$ coefficients, with $K = 1 + \sum_{\sigma=1}^{\sigma}(1 + 2^\sigma)$, instead of $K$ shearlet elements [c.f. Eq. (5)]. As a consequence, the dimensions of the optimization space are reduced, and so are the computation times. Table 1 below summarizes the indices $k$ of the shearlet elements that are restricted per decomposition scale $\sigma$. The set of restricted columns of $\Phi$ then results from applying Eq. (14) with the restricted indices $k$.

By removing the most coherent shearlets, the artificial information –caused by the vertical stripes– is removed from the image, leading to a more incoherent sensing and an improved interpolation when the method is run on a computer.

4. Computer implementation

Experience solving (P_1) with interior point methods such as CVX [45] reveals major computation times. This is partly attributed to the dimensions of the data and the redundancy of the dictionary representation,

---

1The coarsest shearlet $\sigma = 0$ has the lowest resolution [e.g. Fig. 2(a)], and is left unrestricted. A reason not to restrict it is that the sound field at the coarsest scale varies more slowly than the gaps, and it hardly interacts with the mask.
Table 1: Set of indices $k \in [1, K]$ of all shearlets, and of the restricted shearlets, at scales $\sigma = 0$ to 5. Underlined indices denote strictly vertical shearlets. These indices are in correspondence with the indices in the fast finite shearlet transform [41].

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>All indices</th>
<th>Restricted indices</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>1</td>
<td>2–5</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>6–13</td>
<td>6, 7, 13</td>
</tr>
<tr>
<td>3</td>
<td>14–29</td>
<td>14, 15–17, 27–29</td>
</tr>
<tr>
<td>4</td>
<td>30–61</td>
<td>30, 31–37, 55–61</td>
</tr>
<tr>
<td>5</td>
<td>62–125</td>
<td>62, 63–77, 111–125</td>
</tr>
</tbody>
</table>

but mostly due to CVX’s incompatibility with FFTs. A more efficient approach is to run $\nu = 1, \ldots, \nu_{\text{max}}$ Landweber descent iterations, and apply a soft-thresholding operator [32, 46] to each one

$$\alpha^{(\nu)} = T_\zeta \left\{ \alpha^{(\nu-1)} + \hat{\Phi}^H \left[ \hat{y} - \hat{\Phi} \alpha^{(\nu-1)} \right] \right\},$$

also known as the iterative soft-thresholding algorithm (ISTA), where operating a soft-thresholding on the $i$-th entry of a vector $\alpha$ reads

$$T_\zeta \{ \alpha_i \} = \text{sgn} \{ \alpha_i \} \cdot \max \{ 0, |\alpha_i| - \zeta/2 \}.$$  

The interpolation of the RIR image is then carried out at the last iteration via

$$y = \hat{D} \alpha^{(\nu_{\text{max}})}.$$  

4.1. Choosing thresholds

There are various ways of choosing the threshold value, for example the L-curve [52] and ten-fold cross-validation [53]. In all the experiments investigated here, an approach that has provided both efficient and accurate computations is to choose an initial threshold with the L-curve, and thereafter cool this threshold down to the noise level of the measurements. The cooling strategy has been taken from the reference [32].

In mathematical terms, the expression (17) begins with initial solution $\alpha^{(0)} = \hat{\Phi}^H \hat{y}$, i.e. keeping originally measured responses, and the initial threshold is set to

$$\zeta_0(\beta) = \beta \| \alpha^{(0)} \|_\infty,$$

where the maximum (or $\infty$-) norm of $\alpha$ is defined as $\max_i |\alpha_i|$, and $\beta$ is some positive constant that is chosen with the L-curve. Here the L-curve is defined as the logarithmic plot of the sparsity norm $\eta(\beta) = \| \alpha^{(1)}(\beta) \|_1$ versus the residual norm $\rho(\beta) = \| \hat{y} - \hat{\Phi} \alpha^{(1)}(\beta) \|_2$. In other words, given a set of values of $\beta$, $\alpha^{(1)}$ is obtained via Eq. (17) for the set of values of $\zeta_0(\beta)$, and thereafter used to calculate the corresponding pairs $(\eta(\beta), \rho(\beta))$. Then, the curvature function [52]

$$\mathcal{J}(\beta) = \frac{\rho(\beta)'' \eta(\beta)' - \rho(\beta)' \eta(\beta)''}{(\rho(\beta)')^2 + (\eta(\beta)')^2}$$

attains a maximum corresponding to a trade-off between the signal sparsity and the model misfit. Subscripts $t$ and $n$ denote first- and second-order derivatives, which can be approximated numerically via, e.g., cubic splines. Hence, the optimal parameter $\beta^*$ is selected automatically as the solution of $\arg \max_{\beta} \mathcal{J}(\beta)$.

As suggested in [32], the threshold for $\nu \geq 1$ can then be cooled-down - in this work, decreased linearly, via

$$\zeta_\nu = \zeta_0(\beta^*) - \frac{\nu}{\nu_{\text{max}}} (\zeta_0(\beta^*) - \epsilon), \quad \nu = 1, \ldots, \nu_{\text{max}},$$

where $\epsilon$ is the spatially-averaged noise level of the measurements $\hat{y}$. The noise level at each microphone position is estimated once the responses have decayed, using the ITA Toolbox for room acoustics [54].

11
The iterative thresholding algorithm can be summarized as follows. With the aim of promoting sparse solutions, the initial threshold in Eq. (20) is a function of the absolute largest entry of $\alpha^{(0)}$, ideally admitting only the strongest shearlet coefficients at the first iteration. Thereafter, the threshold decreases every iteration so as to gradually increase the $\ell_2$-norm penalty due to model misfit [46]. The number of iterations $\nu_{\text{max}}$ provides a trade-off between accuracy and computation time (more details in Sec. 5.2). Information regarding the computer code of the implementation can be found in Appendix A.

4.2. Leakage bias

Similarly to the case of leakage in Fourier acoustics [55], the lack of spatial periodicity in a RIR image causes shearlets to wraparound and appear as spurious artifacts in the edges of the interpolated image. One counteraction is to pad the spatial dimension with a sound field extrapolation technique. To do this, let us define the number of spatial samples to be appended as $\Delta M = M - M$, here assumed to be even. Then, $\Lambda$ should be rewritten as

$$\Lambda = \begin{pmatrix} I_{\Delta m/2} & 0 \\ 0 & I_{\Delta m/2} \end{pmatrix} \in \mathbb{R}^{(m-p) \times m},$$

(23)

where $I_{\Delta m/2}$ is the identity matrix of size $T \times \Delta M/2$. The matrices $S$, $D$, and $E$ must also be reshaped accordingly. After interpolation the image is cropped back to the original $M$ microphone positions.

5. Experimental study

Experiments have been conducted with linear microphone array responses measured in three rooms, located at the Department of Aeronautical and Vehicle Engineering at KTH, of which photographs are shown in Fig. 9. They are presented from largest to smallest in volume. The source is a small 8Ω loudspeaker, whose sound radiation is omnidirectional [2] between 500 Hz and 4.5 kHz. In all setups, the acoustic center of the source is at a height of 1.5 m (see Fig. 10 below). The speaker is driven with a B&K 2706 amplifier, and the pressure is recorded with a 1/4 in. GRAS 40BD free-field microphone and an NI cDAQ-9178 acquisition system. The microphone positions are at a height of 1.2 m.

The RIRs are obtained by means of logarithmic sweeps in a frequency range$^2$ from 100 Hz to $f_{\text{max}} = 4.5$ kHz. The sweeps are 2 seconds long, which is enough time for the RIRs to completely decay in all the rooms investigated. In order to increase the SNR, the procedure is repeated 10 times per microphone position, such that the final response is the result of averaging the 10 responses. The microphone responses are measured in a linear arrangement of $M = 100$ positions, spaced by $d = 3$ cm. This corresponds to a wavenumber sampling rate$^3$ $\psi_s = [c/(f_{\text{max}}d)] \psi_{\text{max}} \approx 2.5\psi_{\text{max}}$, provided $c = 340$ m/s. Every microphone recording is sampled with rate $f_s = 2.5f_{\text{max}} = 11250$ Hz. In addition, the analysis time window is matched with the reverberation time of the room. This gives RIR images $y_{\text{ref}}$ of size $m = 7092 \times 100$, 5179 \times 100, and 3623 \times 100, respectively, for the three rooms. The total measurement time is on average 2.5 hours per room.

The input RIR images $\hat{y}$ are obtained by under-sampling the microphone positions from 100 to 20 and from 100 to 33, as shown in Fig. 10 together with the source positions in the three rooms. For these two under-sampling factors the RIR images have maximum gaps (between missing microphone positions) of 24 cm and 15 cm respectively. Linear predictive border padding [56] is applied to extrapolate the spatial dimension to $M = 128$ (see Sec. 4.2).

In order to assess the interpolation performance quantitatively, the normalized mean-squared error

$$\text{NMSE} = 10 \log_{10} \frac{1}{M} \sum_{i=1}^{M} \frac{\|\{y\}_{(i,:)} - \{y_{\text{ref}}\}_{(i,:)}\|^2}{\|\{y_{\text{ref}}\}_{(i,:)}\|^2}$$

(24)

$^2$ Frequencies below 500 Hz lie outside the range of the source’s omnidirectional directivity. Therefore, the signal-to-noise ratio may decrease in the frequency range between 100 and 500 Hz.

$^3$ This is equivalent to a target spatial resolution of $c/(2f_{\text{max}}) \approx 3.8$ cm. Then, spatial down-sampling by a factor $u > 2.5/2 = 1.25$ poses the challenge of recovering spatial details smaller than $u \times 3.8$ cm.
is calculated, which is an indicator of the spatially-averaged difference between the interpolated and the reference RIR images. Here \( \cdot \rangle \rangle_{\cdot} \) indexes the vector at the \( i \)-th microphone position and all time instants. The Pearson correlation coefficient

\[
C(f) = \frac{\langle \langle \mathbf{y}(j,:) \rangle \rangle_{\cdot} \cdot \langle \mathbf{y}_{\text{ref}}(j,:) \rangle \rangle_{\cdot}}{\lVert \langle \mathbf{y}(j,:) \rangle \rangle_{\cdot} \rVert \lVert \langle \mathbf{y}_{\text{ref}}(j,:) \rangle \rangle_{\cdot} \rVert}
\]  

(25)
The linear interpolation method used for comparison is based on applying a wavenumber-frequency low-pass filter, containing plane waves inside the support of the far-field plenacoustic function, as indicated by the hourglass-shaped region in Fig. 8. Considering a wavenumber-frequency domain defined by the 2D grid of wavenumbers \( \psi \in [-\psi_s/2, \psi_s/2 - \psi_s/M] \in \mathbb{R}^M \) and frequencies \( f \in [-f_s/2, f_s/2 - f_s/T] \in \mathbb{R}^T \), the hourglass filter is realized in the computer as an ideal truncation filter: assigning amplitude one to propagating waves \( |\psi| \leq 2\pi f/c \), and zero to evanescent waves \( |\psi| > 2\pi f/c \). Then, the linear interpolation entails a 2D FFT of the under-sampled image, followed by a point-wise multiplication with the hourglass filter, and a 2D inverse FFT.

### 5.1. Interpolation performance

Table 2 shows the NMSE for all rooms, using 20 and 33 microphones, and obtained via linear interpolation as well as via thresholding of shearlet dictionaries of various decomposition scales \( \tau \). It can be seen that the NMSE values obtained with linear interpolation are in all cases larger than the NMSE values obtained via thresholding of shearlet coefficients. This can to a considerable extent be attributed to the presence of strong near-field content in the measurements, which lies outside the hourglass filter of the linear interpolation. Another observation is that the NMSE values obtained with shearlets are comparable with those obtained in previous studies [22, 26], with the difference here that scattering elements such as chairs, tables, and fan ducts, as well as sharp edges and uneven surfaces are present in the rooms, and their geometries are not convex. In addition, the NMSE values using shearlets improve as more decomposition scales \( \tau \) are used, i.e. a more redundant dictionary. Despite increasing the computation time, the redundancy of the shearlet representation results in a broader diversity of waves that can be approximated, as well as an increased robustness against missing information [43]. Moreover, the NMSE values increase with more missing information, attributed to the lower probability of recovering the information from a larger proportion of missing data [19].

Table 2: NMSE values obtained using 20 and 33 (out of 100) microphones, via linear interpolation and via thresholding the shearlet coefficients of dictionaries of \( \tau = 2, 3, 4, \) and 5 decomposition scales. Here \( \nu_{\text{max}} = 30 \) iterations. The values of \( \beta^* \) are also included. The lowest NMSE values are bold-faced.

<table>
<thead>
<tr>
<th>Room</th>
<th>Mics</th>
<th>Linear</th>
<th>Shearlets (( \tau = 2 ))</th>
<th>Shearlets (( \tau = 3 ))</th>
<th>Shearlets (( \tau = 4 ))</th>
<th>Shearlets (( \tau = 5 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Munin</td>
<td>20</td>
<td>-1.7 dB</td>
<td>-3 dB</td>
<td>0.043</td>
<td>-4.3 dB</td>
<td>0.037</td>
</tr>
<tr>
<td></td>
<td>33</td>
<td>-3 dB</td>
<td>-6.6 dB</td>
<td>0.035</td>
<td>-8.5 dB</td>
<td>0.026</td>
</tr>
<tr>
<td>Freja</td>
<td>20</td>
<td>-1.7 dB</td>
<td>-2.8 dB</td>
<td>0.07</td>
<td>-3.7 dB</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>33</td>
<td>-3 dB</td>
<td>-5.8 dB</td>
<td>0.053</td>
<td>-7.4 dB</td>
<td>0.049</td>
</tr>
<tr>
<td>Balder</td>
<td>20</td>
<td>-1.6 dB</td>
<td>-2.7 dB</td>
<td>0.053</td>
<td>-3.8 dB</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>33</td>
<td>-2.9 dB</td>
<td>-5.7 dB</td>
<td>0.043</td>
<td>-7.6 dB</td>
<td>0.048</td>
</tr>
</tbody>
</table>

Table 2 also includes the values of \( \beta^* \) obtained from the L-curve at the first iteration of the shearlet thresholding algorithm. These values seem to decrease with increasing number of microphones, as well as increasing number of decomposition scales \( \tau \). What this means is that the \( \ell_1 \)-norm term in (P) is penalized more strongly than the \( \ell_2 \)-norm as the dictionary has fewer scales. As a consequence, the L-curve suggests a (large) value of \( \beta^* \) corresponding to a solution with small \( \ell_2 \)-norm residual, which often results in a smoothening (low-pass filtering) of the gaps. This phenomenon shows up in the interpolated RIR images, particularly in Figs. 11(c), (g), and (k), in which 20 of 100 microphones are used. Another possible reason for said values of \( \beta^* \) is that the gap size is too long compared with the shearlet length, as discussed in Sec. 3.1, resulting additionally in more shearlets of coarser scales (larger sizes) being needed. Nonetheless, these very figures 11(c), (g), and (k) show that the narrower gaps are correctly interpolated (e.g. at \( x \approx 0.5 \) m), and the interpolated responses agree reasonably well with the reference responses. An interesting conclusion is also computed, which is an indicator, ranging from 0 to 1, of the spatial similarity between the room frequency responses at index \( j \). Here \( \tilde{y} \) denotes a temporal FFT applied to every microphone recording, and \( \{ \cdot \}_j \) denotes indexing the vector at the \( j \)-th frequency and all microphone positions.
from this result is that the microphones can be spaced more narrowly in regions of the room where a higher spatial accuracy is desired; and in other cases the spacing can be broadened. Figures 11(b), (f), and (j) demonstrate that linear interpolation introduces strong aliases in the recovered images, which explains the corresponding higher NMSE values in Table 2. Also, the amplitude of the images recovered via linear interpolation is under-estimated (see also Fig. 14 below); thus, for the sake of easier visualization (using the same color scale in all images), the responses in Figs. 11(b), (f), and (j) are actually multiplied by a factor of three.

A comparison between rooms can be drawn at this point. It can be seen in Fig. 11 that the first 15 ms of data are denser in Freja (h) and Balder (l) than in Munin (d). As with Fourier dictionaries [26], a higher density of waves –particularly in the presence of scatterers– leads to a less sparse representation, hence more challenging conditions [19] for a successful recovery using shearlets. This can explain why the NMSE values in Freja and Balder are larger than those in Munin. Later in time, such as in Fig. 12, the responses are dense in the three rooms. The individual differences are partly related to the size of the room: the larger it is, the longer it takes for the sound field to be reflected from the walls. However, before reaching the walls, the sound field encounters scattering and diffracting objects, which may influence the time at which the reflections begin to overlap [57]. In addition, the results in Fig. 12 demonstrate a qualitatively more accurate interpolation due to the increase in microphones from 20 to 33. In particular, Figs. 12(c), (g), and (k) also shows that the interpolation (using shearlets) is accurate in the late part of the responses, which on its own is an improvement with respect to previous studies [21, 22]. As in the case of Fig. 11, the responses in Figs. 12(b), (f) and (j) are multiplied by a factor of three in order to make the visual comparison easier. Although less obviously compared with Fig. 11, the linear interpolation method introduces aliases in the recovered images, justifying the increase in NMSE values with respect to the interpolation using shearlets.

Figure 13 below shows the Pearson correlation coefficient values $C(f)$ for the three rooms, calculated...
Figure 12: RIR interpolation using 33 of 100 microphones, times between 78 and 93.5 ms. Top row (a)-(d): Munin. Middle row (e)-(h): Freja. Bottom row (i)-(l): Balder. Columns read from left to right. First column: under-sampled RIR images. Second column: interpolated RIR images using low-pass filter. Third column: interpolated RIR images using shearlets, with $\tau = 5$ and $\nu_{\text{max}} = 30$. Fourth column: reference RIR images.

from the responses obtained via linear interpolation and via shearlet-based thresholding, using 20 and 33 microphones. In all cases, the $C$ values obtained with shearlet interpolation are closer to 1 in a broader bandwidth compared to those obtained with the linear interpolation; which further justifies the differences in accuracy between the two methods. It can also be seen that an increase in number of microphones (less information missing) leads to $C$ values that are closer to 1 in a broader bandwidth. In particular, for frequencies below 500 Hz, the $C$ values are potentially affected by the lack of omnidirectivity of the source.

On another note, the vertical dashed lines in Fig. 13 show the aliasing-free (AF) rates that result from under-sampling the data: 750 and 1875 Hz for 20 and 33 microphones, respectively. When making use of shearlets, the $C$ values in Fig. 13 are close to 1 between 500 Hz and the AF rates, indicating that aliasing is absent and that the interpolation is accurate. As the frequency increases beyond the AF rates, however, the $C$ values decrease. Using 20 microphones in Munin and Balder leads to $C$ values between 0.6 and 1 in a bandwidth of nearly twice the AF rate, which indicates that the responses interpolated with shearlets are not far from the reference responses in such bandwidth. A similar result is obtained using 33 microphones in these two rooms. In Freja, however, the $C$ values using 20 (33) microphones drop down to 0.1 (0.4) around 3000 Hz, which is attributed to a more complex wave interference [e.g. compare Figs. 11(h) with (d) and (l)].

Such interference in Freja can be seen in the neighborhood of 3000 Hz in Fig. 14, in the form of dense resonances. This can be a challenge for a successful interpolation, since there is a greater loss of sparsity due to the presence of more waves. On the one hand, Fig. 14 shows that linear interpolation under-estimates the responses by a considerable amount of decibels in nearly the whole frequency range. On the other hand, the interpolated responses using shearlets is rather inaccurate below 500 Hz, and it is most accurate between 500 Hz and slightly above the AF rate. At this particular point in space ($x = 6$ cm), the gap is 15 cm and the responses interpolated using shearlets agree well with the reference ones. However, visual inspection of
Figure 13: Pearson correlation coefficients in the frequency range between 100 and 4500 Hz, using linear interpolation (dark grey circles) and using shearlets (green solid lines). Top row (a)-(b): Munin. Middle row (c)-(d): Freja. Bottom row (e)-(f): Balder. Left column: using 33 out of 100 microphones. Right column: using 20 out of 100 microphones. Here $\tau = 5$, and $\nu_{\text{max}} = 30$. Vertical dashed lines denote the aliasing-free (AF) rate after under-sampling.

Figure 14: Normalized sound pressure levels (SPL) versus frequency at array position $x = 6$ cm in (a) Munin, (b) Freja, and (c) Balder. Solid black: measured (reference) SPL responses. Solid green: recovered SPL responses using shearlet-based interpolation, with $\tau = 5$, and $\nu_{\text{max}} = 30$. Dash-dotted blue: recovered SPL using linear interpolation. The interpolation is done using 20 out of 100 microphones.

Figs. 11(c), (g), and (k) suggests that this agreement degrades for responses interpolated within the widest gaps (e.g. the 24 cm gap centered at $x \approx 0.3$ m).

A fair amount of modes is accurately resolved using shearlets, even well above the AF rates [e.g. Fig. 14 at $f = 2150$ Hz]. In the higher frequency end the interpolated responses are slightly under-estimated, possibly due to the smoothening (low-pass filtering effect) of the gaps mentioned earlier. More evidence of such an effect can be found in Fig. 15, which shows that the RIRs interpolated using shearlets are slightly under-estimated in the three rooms. In particular, the interpolation accuracy seems to drop as time progresses, which can be associated with the fact that faster variations increasingly dominate the sound field.
as time progresses [57].

Figure 15: Measured RIRs (black) and interpolated RIRs using shearlets (yellow) at array position $x = 6$ cm, obtained using 20 of 100 microphones, provided $\tau = 5$ and $\nu_{\text{max}} = 30$. (a) Lecture room Munin. (b) Lecture room Freja. (c) Meeting room Balder. The subplot in the top-right of every figure is a zoom-in of the first 100 ms of the responses.

In truth, the difficulty of interpolating arbitrarily long responses is that the sound field enriches with faster variations (due to stronger wave interactions) of decreasing signal-to-noise ratio (due to absorption). Ultimately, at some point in time, the responses reach the noise level and it becomes virtually impossible to interpolate the wavefronts. One could conjecture that the proposed method can, in principle, interpolate responses of unrestricted length, as long as the measured signals are well above the noise level [37] (see Sec. 5.3 for more details). However, this statement should be taken with caution, since longer responses lead to longer computation times.

5.2. Computational aspects

Figure 16 shows the NMSE obtained with the proposed method, plotted against number of iterations $\nu_{\text{max}}$, given $\tau = 3$, and using 33 out of 100 microphones. As it can be seen for Munin and Freja, the NMSE stops decreasing when running more than 50 iterations. In Balder this happens with a larger number of iterations $\approx 80$. These differences are partly attributed to the different properties of the rooms, which result in RIR images with different degrees of sparsity; but also possibly due to the different noise levels in the rooms (see Sec. 5.3). The curves also indicate that the NMSE values—and potentially the $C$ values and the qualitative performance of the proposed method itself, improve further (compared with the results shown in Sec. 5.1) when the thresholding algorithm runs more iterations.

In addition, the right axis of Fig. 16 shows the estimated computation time (CT). (The values are estimated in a 2.7 GHz i7, 16 GB RAM portable computer.) The different CTs between rooms is due to the different image sizes. It should be mentioned that the linear interpolation method is considerably faster (in the order of seconds), but it clearly fails at providing accurate results. Coming back to the challenge in Sec. 1 of acquiring meaningful enough information with limited resources, the proposed method points towards a more suitable means of compromising between the spatial resolution and the effort it takes to acquire (and interpolate) RIRs in rooms with unknown properties and contents.

5.3. Remarks on signal-to-noise ratios

As the measured RIRs have a rather high signal-to-noise ratio (SNR)—in the order of 100 dB in Munin and Balder, and 60 dB in Freja, this section includes a complementary performance evaluation of the proposed
method against lower values of SNR. To this end, Gaussian noise is added to the reference RIRs, provided SNR values of 20, 30, 40, and 50 dB. The noisy, under-sampled responses are interpolated with the proposed method, using a dictionary of \( \tau = 3 \) scales, and running \( \nu_{\max} = 30 \) thresholding iterations. The NMSE values are calculated from the responses recovered in the three rooms, using 20 and 33 microphones. Figure 17 summarizes these results.

As these curves show, the NMSE increases with decreasing SNR, as one would expect. In addition, the NMSE values at a SNR below 30 dB collapse at approximately the same value (NMSE \( \approx -0.5 \) dB at SNR = 20), independently of the under-sampling factor \( u \). What this suggests is that the accuracy of the recovery is most influenced by the presence of background noise, rather than by the lack of information. As the SNR increases beyond 30 dB, the NMSE values obtained by using 33 and 20 microphones begin to differ, which shows that the accuracy of the recovery is most influenced by the difference in missing information.

On another note, the use of the L-curve to obtain the threshold parameter may be inadequate when the SNR is too low (e.g. below 30 dB), and another parameter selection method should be adopted, for instance the ten-fold cross-validation [53]. It should also be stressed that the ISO standards [1–3] do not recommend the measurement of RIRs under too low SNR conditions. Likewise, it would not be fair to recommend applying the proposed method to measurements performed in too noisy environments.

6. Conclusions

The present paper introduces a method to recover unmeasured room impulse responses (RIRs) in acoustical spaces with unknown properties and contents, by means of compressive sensing. The main contribution
of this work is to enable the recovery of unmeasured RIRs in more complex (realistic) acoustical spaces, while keeping the number of microphones to a minimum. The central idea is to sparsely represent the RIRs with shearlets [33], which are curved elementary functions parametrized with various sizes, orientations, and translations; and can be understood as sparse approximations of the acoustic wavefronts in the room. The recovery of the missing RIRs is formulated as an interpolation task, in the form of an $\ell_1$-norm optimization problem, and the solution is found via iterative thresholding of the shearlet coefficients.

Experiments in three different rooms are conducted with linear (1D) microphone array measurements in a frequency bandwidth between 100 and 4500 Hz, with the aim of examining the performance of the proposed method. For the sake of comparison, the responses are also linearly interpolated using a low-pass filter in the wavenumber-frequency domain. The results demonstrate that linear interpolation is significantly outperformed by the proposed method in the three rooms. In all the cases investigated, the responses recovered using shearlets agree reasonably well with the reference responses, even in the presence of scattering and diffracting elements, and in rooms with non-convex geometries and unknown damping; none of which has been investigated by the methods proposed in the existing literature. It can be conjectured that secondary wavefronts, potentially originating from diffraction events [58], may also be sparse in shearlet dictionaries. Overall, the accuracy of the proposed method improves with more dictionary scales, and more thresholding iterations; and it deteriorates when less information is available (more of the faster variations of the sound field become unavailable). In addition, the proposed method is not recommended for measurements performed in environments that are too noisy (too low signal-to-noise ratio conditions). It is also shown that the computation time to recover the missing responses is very reasonable: in the order of a couple of minutes for RIR data as big as 700 thousand samples in the largest room investigated. This is all the more impactful compared with the time it took to measure all the responses in that room, which exceeded 2 hours.

Research intentions following this work concern the extension and validation of the method in higher spatial dimensions, which as of today is partly limited by: 1) the dimensions of the available computer implementations of shearlet dictionaries (typically tailored for image, 2D, and video, 3D, signals); and 2) the dimensions of RIR data and the associated computational complexity and storage required to execute the optimization routines. As a last remark, shearlet decompositions of RIR measurements can be regarded as representations of the sound field at different wavelengths, capturing various wave directions and times of arrival. It can be the case that these representations are also useful in applications such as sound source characterization [55], sound field separation [59], measurement of absorption [60]; or in cases in which a multi-resolution analysis of the acoustic field is desirable.

7. Acknowledgments

The author would like to acknowledge the anonymous reviewers for their valuable comments and suggestions. This work is financially supported by the Swedish Research Council (Vetenskapsrådet), under grant agreement No. 2016-04366. Thanks to Prof. I. Lopez Arteaga for the steady advice and support, and to Prof. G. Kutyniok for the references and theoretical aspects of shearlet inpainting. Warm thanks to R. Hamden, C. O’Reilly, M. Laudato, J. Synnelius, and S. Chatterjee for the valuable comments and suggestions.

Appendix A. Reproducible research and pseudocodes

Having the thought of reproducible research [61] in mind, the following link https://kth.box.com/shared/static/rr7ev4vqg3au3hpchzscwrfhgb7b5r4.zip allows the reader to download Matlab code to reproduce the results in Figs. 11-17, as well as the RIR data measured in the three rooms. In addition, the following pseudocode in Algorithm 1 summarizes the computer implementation of the proposed method. The inputs are: the under-sampled RIR image $\hat{y}$, the number of shearlet decomposition scales $\tau$, and the number of thresholding iterations $\nu_{\text{max}}$. The fast finite shearlet transform [41] is used to carry out the operations performed by $D$ and $S$. 

20
Algorithm 1: interpolate_RIRs

1: Initialize $\alpha^{(0)} \leftarrow \hat{\Phi}^H \hat{y}$
2: Define set of threshold parameters $\beta = \{\beta_1, ..., \beta_B\}$
3: for $i = 1, ..., B$ do
4:   Set $\zeta(\beta_i) = \beta_i ||\alpha^{(0)}||_\infty$
5:   Threshold $\alpha(\beta_i) \leftarrow T_{\zeta(\beta_i)}\{\alpha^{(0)} + \hat{\Phi}^H \left[\hat{y} - \hat{\Phi} \alpha^{(0)}\right]\}$
6:   Compute $\eta(\beta_i) = ||\alpha(\beta_i)||_1$
7:   Compute $\rho(\beta_i) = ||\hat{y} - \hat{\Phi} \alpha(\beta_i)||_2$
8:   Approximate $\eta(\beta)', \eta(\beta)''$, $\rho(\beta)'$, and $\rho(\beta)''$ with cubic splines
9:   Solve $\beta^* = \arg \max_{\beta} J(\beta)$ [see Eq. (21)]
10: Set $\zeta_0 \leftarrow \beta^* ||\alpha^{(0)}||_\infty$
11: Threshold $\alpha^{(1)} \leftarrow T_{\zeta_0}\{\alpha^{(0)} + \hat{\Phi}^H \left[\hat{y} - \hat{\Phi} \alpha^{(0)}\right]\}$
12: Set $\nu \leftarrow 0$
13: while $||\hat{y} - \hat{\Phi} \alpha^{(\nu)}||_2 > \epsilon$, or $\nu < \nu_{\text{max}}$ do
14:   Update $\nu \leftarrow \nu + 1$
15:   Update $\zeta \leftarrow \zeta_0 - \frac{\nu}{\nu_{\text{max}}} (\zeta_0 - \epsilon)$
16:   Threshold $\alpha^{(\nu)} \leftarrow T_{\zeta}\{\alpha^{(\nu-1)} + \hat{\Phi}^H \left[\hat{y} - \hat{\Phi} \alpha^{(\nu-1)}\right]\}$
17: Interpolate $\hat{y} \leftarrow \hat{D} \alpha^{(\nu_{\text{max}})}$

References


