Acoustic echoes reveal room shape

Ivan Dokmanić,† Reza Parhizkar, Andreas Walther, Yue M. Lu, and Martin Vetterli

*Audiovisual Communications Laboratory (LCAV), School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; and †Signals, Information, and Networks Group (SING), School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved May 17, 2013 (received for review January 4, 2013)

Imagine that you are blindfolded inside an unknown room. You snap your fingers and listen to the room’s response. Can you hear the shape of the room? Some people can do it naturally, but can we design computer algorithms that hear rooms? We show how to compute the shape of a convex polyhedral room from its response to a known sound, recorded by a few microphones. Geometric relationships between the arrival times of echoes enable us to “blindfoldedly” estimate the room geometry. This is achieved by exploiting the properties of Euclidean distance matrices. Furthermore, we show that under mild conditions, first-order echoes provide a unique description of convex polyhedral rooms. Our algorithm starts from the recorded impulse responses and proceeds by learning the correct assignment of echoes to walls. In contrast to earlier methods, the proposed algorithm reconstructs the full 3D geometry of the room from a single sound emission, and with an arbitrary geometry of the microphone array. As long as the microphones can hear the echoes, we can position them as we want. Besides answering a basic question about the inverse problem of room acoustics, our results find applications in areas such as architectural acoustics, indoor localization, virtual reality, and audio forensics.

I. Introduction

A famous question asks whether one can hear the shape of a drum? More concretely, does the shape of a drum depend on its resonant modes (eigenvectors)? This problem is related to a question in astrophysics (2), and the answer turns out to be negative: Using tools from group representation theory, Gordon et al. (3, 4) presented several elegantly constructed counterexamples, including the two polygonal membranes of different shapes necessarily resonate at different frequencies. Still, these drums would hardly sound the same if hit with a drumstick. They share the resonant frequencies, but the impulse responses are different. Even a single drum struck at different points sounds differently. In this work, we ask a similar question about rooms. Assume you are blindfolded inside a room; you snap your fingers and listen to echoes. Can you hear the shape of the room? Intuitively, and for simple room shapes, we know that this is possible. A shoebox room, for example, has well-defined modes, from which we can derive its size. However, the question is challenging in more general cases, even if we presume that the room impulse response (RIR) contains an arbitrarily long set of echoes (assuming an ideal, noiseless measurement) that should specify the room geometry.

It may appear that Kac’s question and the problem we pose are equivalent. This is not the case, for the sound of a drum depends on more than just its resonant frequencies (eigenvalues)—it also depends on its resonant modes (eigenvectors). In the paper “Drums that sound the same” (5), Chapman explains how to construct drums of different shapes with matching resonant frequencies. Still, these drums would hardly sound the same if hit with a drumstick. They share the resonant frequencies, but the impulse responses are different. Even a single drum struck at different points sounds differently. In this work, we ask a similar question about rooms. Assume you are blindfolded inside a room; you snap your fingers and listen to echoes. Can you hear the shape of the room? Intuitively, and for simple room shapes, we know that this is possible. A shoebox room, for example, has well-defined modes, from which we can derive its size. However, the question is challenging in more general cases, even if we presume that the room impulse response (RIR) contains an arbitrarily long set of echoes (assuming an ideal, noiseless measurement) that should specify the room geometry.

It may appear that Kac’s question and the problem we pose are equivalent. This is not the case, for the sound of a drum depends on more than just its resonant frequencies (eigenvalues)—it also depends on its resonant modes (eigenvectors). In the paper “Drums that sound the same” (5), Chapman explains how to construct drums of different shapes with matching resonant frequencies. Still, these drums would hardly sound the same if hit with a drumstick. They share the resonant frequencies, but the impulse responses are different. Even a single drum struck at different points sounds differently. In this work, we ask a similar question about rooms. Assume you are blindfolded inside a room; you snap your fingers and listen to echoes. Can you hear the shape of the room? Intuitively, and for simple room shapes, we know that this is possible. A shoebox room, for example, has well-defined modes, from which we can derive its size. However, the question is challenging in more general cases, even if we presume that the room impulse response (RIR) contains an arbitrarily long set of echoes (assuming an ideal, noiseless measurement) that should specify the room geometry.

In this work, we ask a similar question about rooms. Assume you are blindfolded inside a room; you snap your fingers and listen to echoes. Can you hear the shape of the room? Intuitively, and for simple room shapes, we know that this is possible. A shoebox room, for example, has well-defined modes, from which we can derive its size. However, the question is challenging in more general cases, even if we presume that the room impulse response (RIR) contains an arbitrarily long set of echoes (assuming an ideal, noiseless measurement) that should specify the room geometry.

In this work, we ask a similar question about rooms. Assume you are blindfolded inside a room; you snap your fingers and listen to echoes. Can you hear the shape of the room? Intuitively, and for simple room shapes, we know that this is possible. A shoebox room, for example, has well-defined modes, from which we can derive its size. However, the question is challenging in more general cases, even if we presume that the room impulse response (RIR) contains an arbitrarily long set of echoes (assuming an ideal, noiseless measurement) that should specify the room geometry.

In this work, we ask a similar question about rooms. Assume you are blindfolded inside a room; you snap your fingers and listen to echoes. Can you hear the shape of the room? Intuitively, and for simple room shapes, we know that this is possible. A shoebox room, for example, has well-defined modes, from which we can derive its size. However, the question is challenging in more general cases, even if we presume that the room impulse response (RIR) contains an arbitrarily long set of echoes (assuming an ideal, noiseless measurement) that should specify the room geometry.

In this work, we ask a similar question about rooms. Assume you are blindfolded inside a room; you snap your fingers and listen to echoes. Can you hear the shape of the room? Intuitively, and for simple room shapes, we know that this is possible. A shoebox room, for example, has well-defined modes, from which we can derive its size. However, the question is challenging in more general cases, even if we presume that the room impulse response (RIR) contains an arbitrarily long set of echoes (assuming an ideal, noiseless measurement) that should specify the room geometry.

In this work, we ask a similar question about rooms. Assume you are blindfolded inside a room; you snap your fingers and listen to echoes. Can you hear the shape of the room? Intuitively, and for simple room shapes, we know that this is possible. A shoebox room, for example, has well-defined modes, from which we can derive its size. However, the question is challenging in more general cases, even if we presume that the room impulse response (RIR) contains an arbitrarily long set of echoes (assuming an ideal, noiseless measurement) that should specify the room geometry.

In this work, we ask a similar question about rooms. Assume you are blindfolded inside a room; you snap your fingers and listen to echoes. Can you hear the shape of the room? Intuitively, and for simple room shapes, we know that this is possible. A shoebox room, for example, has well-defined modes, from which we can derive its size. However, the question is challenging in more general cases, even if we presume that the room impulse response (RIR) contains an arbitrarily long set of echoes (assuming an ideal, noiseless measurement) that should specify the room geometry.

In this work, we ask a similar question about rooms. Assume you are blindfolded inside a room; you snap your fingers and listen to echoes. Can you hear the shape of the room? Intuitively, and for simple room shapes, we know that this is possible. A shoebox room, for example, has well-defined modes, from which we can derive its size. However, the question is challenging in more general cases, even if we presume that the room impulse response (RIR) contains an arbitrarily long set of echoes (assuming an ideal, noiseless measurement) that should specify the room geometry.

In this work, we ask a similar question about rooms. Assume you are blindfolded inside a room; you snap your fingers and listen to echoes. Can you hear the shape of the room? Intuitively, and for simple room shapes, we know that this is possible. A shoebox room, for example, has well-defined modes, from which we can derive its size. However, the question is challenging in more general cases, even if we presume that the room impulse response (RIR) contains an arbitrarily long set of echoes (assuming an ideal, noiseless measurement) that should specify the room geometry.

In this work, we ask a similar question about rooms. Assume you are blindfolded inside a room; you snap your fingers and listen to echoes. Can you hear the shape of the room? Intuitively, and for simple room shapes, we know that this is possible. A shoebox room, for example, has well-defined modes, from which we can derive its size. However, the question is challenging in more general cases, even if we presume that the room impulse response (RIR) contains an arbitrarily long set of echoes (assuming an ideal, noiseless measurement) that should specify the room geometry.

In this work, we ask a similar question about rooms. Assume you are blindfolded inside a room; you snap your fingers and listen to echoes. Can you hear the shape of the room? Intuitively, and for simple room shapes, we know that this is possible. A shoebox room, for example, has well-defined modes, from which we can derive its size. However, the question is challenging in more general cases, even if we presume that the room impulse response (RIR) contains an arbitrarily long set of echoes (assuming an ideal, noiseless measurement) that should specify the room geometry.

In this work, we ask a similar question about rooms. Assume you are blindfolded inside a room; you snap your fingers and listen to echoes. Can you hear the shape of the room? Intuitively, and for simple room shapes, we know that this is possible. A shoebox room, for example, has well-defined modes, from which we can derive its size. However, the question is challenging in more general cases, even if we presume that the room impulse response (RIR) contains an arbitrarily long set of echoes (assuming an ideal, noiseless measurement) that should specify the room geometry.

In this work, we ask a similar question about rooms. Assume you are blindfolded inside a room; you snap your fingers and listen to echoes. Can you hear the shape of the room? Intuitively, and for simple room shapes, we know that this is possible. A shoebox room, for example, has well-defined modes, from which we can derive its size. However, the question is challenging in more general cases, even if we presume that the room impulse response (RIR) contains an arbitrarily long set of echoes (assuming an ideal, noiseless measurement) that should specify the room geometry.

In this work, we ask a similar question about rooms. Assume you are blindfolded inside a room; you snap your fingers and listen to echoes. Can you hear the shape of the room? Intuitively, and for simple room shapes, we know that this is possible. A shoebox room, for example, has well-defined modes, from which we can derive its size. However, the question is challenging in more general cases, even if we presume that the room impulse response (RIR) contains an arbitrarily long set of echoes (assuming an ideal, noiseless measurement) that should specify the room geometry.
Microphones hear the convolution of the emitted sound with the corresponding RIR, \( y_m = x * h_m = \int x(s) h_m(-s) ds \). By measuring the impulse responses we access the propagation times \( \tau_{mi} \), and these can be linked to the room geometry by the image source (IS) model \((17, 18)\). According to the IS model, we can replace reflections by virtual sources. As illustrated in Fig. 2, virtual sources are mirror images of the true sources across the corresponding reflecting walls. From the figure, the image \( \tilde{s}_i \) of the source \( s \) with respect to the \( i \)th wall is computed as

\[
\tilde{s}_i = s + 2(p_i - s, n_i)n_i,
\]

where \( n_i \) is the unit normal, and \( p_i \) any point belonging to the \( i \)th wall. The time of arrival (TOA) of the echo from the \( i \)th wall is \( t_i = |s_i - r_i|/c \), where \( c \) is the speed of sound.

In a convex room with a known source, knowing the image sources is equivalent to knowing the walls—we can search for points instead of searching for walls. The challenge is that the distances are unlabeled: It might happen that the \( k \)th peak in the RIR from microphone 1 and the \( k \)th peak in the RIR from microphone 2 come from different walls. This is illustrated in Figs. 3 and 4. Thus, we have to address the problem of echo labeling. The loudspeaker position need not be known. We can estimate it from the direct sound using either TOA measurements, or differences of TOAs if the loudspeaker is not synchronized with the microphones \((19–21)\).

In practice, having a method to find good combinations of echoes is far more important than only sorting correctly selected echoes. Impulse responses contain peaks that do not correspond to any wall. These spurious peaks can be introduced by noise, nonlinearities, and other imperfections in the measurement system. We find that a good strategy is to select a number of peaks greater than the number of walls and then to prune the selection. Furthermore, some second-order echoes might arrive before some first-order ones. The image sources corresponding to second-order or higher-order echoes \((e.g., \text{Fig. 2})\) will be estimated as any other image source. However, because we can express a second-order image source in terms of the first-order ones as

\[
\tilde{s}_j = \tilde{s}_i + 2(p_j - \tilde{s}_i, n_j)n_j,
\]

we can eliminate it during postprocessing by testing the above two expressions.

**Echo Labeling**

The purpose of echo labeling is twofold. First, it serves to remove the “ghost” echoes (that do not correspond to walls) detected at the peak-picking stage. Second, it determines the correct assignment between the remaining echoes and the walls. We propose two methods for recognizing correct echo combinations. The first one is based on the properties of Euclidean distance matrices (EDM), and the second one on a simple linear subspace condition.

**EDM-Based Approach.** Consider a room with a loudspeaker and an array of \( M \) microphones positioned so that they hear the first-order echoes \((\text{we typically use } M = 5)\). Denote the receiver positions by \( r_1, \ldots, r_M \), \( r_m \in \mathbb{R}^3 \) and the source position by \( s \in \mathbb{R}^3 \). The described setup is illustrated in Fig. 5. We explain the EDM-based echo sorting with reference to this figure. Let \( \mathbf{D} \in \mathbb{R}^{M \times M} \) be a matrix whose entries are squared distances between microphones, \( \mathbf{D}[i,j] = |r_i - r_j|^2 \), \( 1 \leq i, j \leq M \). Here, \( \mathbf{D} \) is an EDM corresponding to the microphone setup. It is symmetric with a zero diagonal and positive off-diagonal entries.

If the loudspeaker emits a sound, each microphone receives the direct sound and \( K \) first-order echoes corresponding to the \( K \) walls. The arrival times of the received echoes are proportional to the distances between image sources and microphones.
already discussed, we face a labeling problem as we do not know which wall generated which echo. This problem is illustrated in Fig. 3 for two walls and in Fig. 4 for the whole room. Simple heuristics, such as grouping the closest pulses or using the ordinal number of a pulse, have limited applicability, especially with larger distances between microphones. That these criteria fail is evident from Fig. 4.

We propose a solution based on the properties of EDMs. The loudspeaker and the microphones are—to a good approximation—points in space, so their pairwise distances form an EDM. We can exploit the rank property: An EDM corresponding to a point set in $\mathbb{R}^n$ has rank at most $(n+2)$ (22). Thus, in 3D, its rank is at most 5. We start from a known point set (the microphones) and want to add another point—an image source. This requires adding a row and a column to $D$, listing squared distances between the microphones and the image source. We extract the list of candidate distances from the RIRs, but some of them might not correspond to an image source; and for those that do correspond, we do not know to which one. Consider again the setup in Fig. 5. Microphone 1 hears echoes from all of the walls, and we augment $D$ by choosing different echo combinations. Two possible augmentations are shown. Here, $D_{aug,1}$ is a plausible augmentation of $D$ because all of the distances correspond to a single image source, and they appear in the correct order. This matrix passes the rank test, or more specifically, it is an EDM. The second matrix, $D_{aug,2}$, is a result of an incorrect echo assignment, as it contains entries coming from different walls. A priori, we cannot tell whether the red echo comes from wall 1 or from wall 2. It is simply an unlabeled peak in the RIR recorded by microphone 1. However, the augmented matrix $D_{aug,2}$ does not pass the rank test, so we conclude that the corresponding combination of echoes is not correct.

To summarize, wrong assignments lead to augmentations of $D$ that are not EDMs. In particular, these augmentations do not have the correct rank. As it is very unlikely (as will be made precise later) for incorrect combinations of echoes to form an EDM, we have designed a tool to detect correct echo combinations.

More formally, let $e_m$ list the candidate distances computed from the RIR recorded by the $m$th microphone. We proceed by augmenting the matrix $D$ with a combination of $M$ unlabeled squared distances $d_{(i_1,\ldots,i_6)}$ to get $D_{aug}$:

$$D_{aug}(d_{(i_1,\ldots,i_6)}) = \begin{bmatrix} D & d_{(i_1,\ldots,i_6)} \\ d_{(i_1,\ldots,i_6)}^\top & 0 \end{bmatrix}.$$  [5]

The column vector $d_{(i_1,\ldots,i_6)}$ is constructed as

$$d_{(i_1,\ldots,i_6)}[m] = e_m[\tilde{d}_m],$$  [6]

with $\tilde{d}_m \in \{1,\ldots,\text{length}(e_m)\}$. In words, we construct a candidate combination of echoes $d$ by selecting one echo from each microphone. Note that length$(e_m) \neq$ length$(e_p)$ for $m \neq n$ in general. That is, we can pick a different number of echoes from different microphones. We interpret $D_{aug}$ as an object encoding a particular selection of echoes $d$.

One might think of EDM as a mold. It is very much like Cinderella’s glass slipper: If you can snugly fit a tuple of echoes in it, then they must be the right echoes. This is the key observation: If $\text{rank}(D_{aug}) < 6$ or more specifically $D_{aug}$ verifies the EDM property, then the selected combination of echoes corresponds to an image source, or equivalently to a wall. Even if this approach requires testing all of the echo combinations, in practical cases the number of combinations is small enough that this does not present a problem.

Subspace-Based Approach. An alternative method to obtain correct echo combinations is based on a simple linear condition. Note that we can always choose the origin of the coordinate system so that

$$\sum_{m=1}^{M} r_m = 0.$$  [7]

Let $\hat{s}_k$ be the location vector of the image source with respect to wall $k$. Then, up to a permutation, we receive at the $m$th microphone the squared distance information,

$$y_{k,m} \overset{\text{def}}{=} ||s_k - r_m||^2 = ||s_k||^2 - 2 \text{vec}(s_k)^\top r_m + ||r_m||^2.$$  [8]

Define further $\bar{y}_{k,m} \overset{\text{def}}{=} \frac{1}{2} (y_{k,m} - ||r_m||^2) = r_m^\top \hat{s}_k - \frac{1}{2}||\hat{s}_k||^2$. We have in vector form

$$\begin{bmatrix} \bar{y}_{k,1} \\ \bar{y}_{k,2} \\ \vdots \\ \bar{y}_{k,M} \end{bmatrix} = \begin{bmatrix} r_1^\top \\ r_2^\top \\ \vdots \\ r_M^\top \end{bmatrix} \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \\ \vdots \\ -\frac{1}{2} \end{bmatrix} ||\hat{s}_k||^2,$$

or $\bar{y}_k = \mathbf{R}\hat{s}_k$.  [9]

Thanks to the condition 7, we have that

$$\sum_{m=1}^{M} r_m = 0.$$
As discussed, in the presence of noise between the points in a room in which our algorithm can be applied. From the range criterion, as it performs well in practice, it is sufficient to verify the linear constraint

\[ \mathbf{1}^T \mathbf{y}_k = -\frac{M}{2} \left\| \mathbf{s}_k \right\|^2 \quad \text{or} \quad \left\| \mathbf{s}_k \right\|^2 = -\frac{2}{M} \sum_{m=1}^{M} \mathbf{y}_{k,m}. \]  

\[ \mathbf{1}^T \mathbf{y}_k = -\frac{M}{2} \left\| \mathbf{s}_k \right\|^2 \quad \text{or} \quad \left\| \mathbf{s}_k \right\|^2 = -\frac{2}{M} \sum_{m=1}^{M} \mathbf{y}_{k,m}. \]  

The image source is found as

\[ \mathbf{s}_k = \mathbf{S} \mathbf{y}_k, \]

where \( \mathbf{S} \) is a matrix satisfying

\[ \mathbf{S} \mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \]

These two conditions characterize the distance information. In practice, it is sufficient to verify the linear constraint

\[ \mathbf{y}_k \in \text{range}(\mathbf{R}), \]

where \( \text{range}(\mathbf{R}) \) is a proper subspace when \( M \geq 5 \). However, note that we can use the nonlinear condition \[10\] even if \( M = 4 \).

**Uniqueness.** Can we guarantee that only one room corresponds to the collected first-order echoes? To answer this, first we define the set of “good” rooms in which our algorithm can be applied. The algorithm relies on the knowledge of first-order echoes, so we require that the microphones hear them. This defines a good room, which is in fact a combination of the room geometry and the microphone array/loudspeaker location.

**Definition 1:** (Feasibility). Given a room \( \mathcal{R} \) and a loudspeaker position \( s \), we say that the point \( x \in \mathcal{R} \) is feasible if a microphone placed at \( x \) receives all the first-order echoes of a pulse emitted from \( s \).

Our argument is probabilistic: The set of vectors \( \mathbf{d} \) such that rank \( \mathbf{D}_{\text{aug}} = 5 \) has measure zero in \( \mathbb{R}^5 \). Analogously, in the subspace formulation, range(\( \mathbf{R} \)) is a proper subspace of \( \mathbb{R}^5 \) thus having measure zero. To use four microphones, observe that the same is true for the set of vectors satisfying \[10\] in \( \mathbb{R}^4 \). These observations, along with some technical details, enable us to state the uniqueness result.\(^*\)

**Theorem 1.** Consider a room with a loudspeaker and \( M \geq 4 \) microphones placed uniformly at random inside the feasible region. Then the unlabeled set of first-order echoes uniquely specifies the room with probability 1. In other words, almost surely exactly one assignment of first-order echoes to walls describes a room.

This means that we can reconstruct any convex polyhedral room if the microphones are in the feasible region. A similar result could be stated by randomizing the room instead of the microphone setup, but that would require us to go through the inconvenience of generating a random convex room. In the following, we concentrate on the EDM criterion, as it performs better in experiments.

**Practical Algorithm**

In practice, we face different sources of uncertainty. One such source is the way we measure the distances between microphones. We can try to reduce this error by calibrating the array, but we find the proposed schemes to be very stable with respect to uncertainties in array calibration. Additional sources of error are the finite sampling rate and the limited precision of peak-picking algorithms. These are partly caused by unknown loudspeaker and microphone impulse responses, and general imperfections in RIR measurements. They can be mitigated with higher sampling frequencies and more sophisticated time-of-arrival estimation algorithms. At any rate, testing the rank of \( \mathbf{D}_{\text{aug}} \) is not a way to go in the presence of measurement uncertainties. The solution is to measure how close \( \mathbf{D}_{\text{aug}} \) is to an EDM. We can consider different constructions:

(i) Heuristics based on the singular values of \( \mathbf{D}_{\text{aug}} \);
(ii) distance of \( \mathbf{y}_k \) from range(\( \mathbf{R} \)) (Eq. 13);
(iii) nonlinear norm condition \[10\]; and
(iv) distance between \( \mathbf{D}_{\text{aug}} \) and the closest EDM.

The approach based on the singular values of \( \mathbf{D}_{\text{aug}} \) captures only the rank requirement on the matrix. However, the requirement that \( \mathbf{D}_{\text{aug}} \) be an EDM brings in many additional subtle dependencies between its elements. For instance, we have that

\[ \mathbf{D}_{\text{aug}} \in \text{EDM} \iff (\mathbf{I} - \frac{1}{M+1} \mathbf{I}) \mathbf{D}_{\text{aug}} (\mathbf{I} - \frac{1}{M+1} \mathbf{I})^\top \preceq 0. \]

Unfortunately \[14\], does not allow us to specify the ambient dimension of the point set. Imposing this constraint leads to even more dependencies between the matrix elements, and the resulting set of matrices is no longer a cone (it is actually not convex anymore). Nevertheless, we can apply the family of algorithms used in multidimensional scaling (MDS) \[23\] to find the closest EDM between the points in a fixed ambient dimension.

**Multidimensional Scaling.** As discussed, in the presence of noise the rank test on \( \mathbf{D}_{\text{aug}} \) is inadequate. A good way of dealing with this nuisance (as verified through experiments) is to measure how close \( \mathbf{D}_{\text{aug}} \) is to an EDM. To this end we use MDS to construct the point set in a given dimension (3D) that produces the EDM “closest” to \( \mathbf{D}_{\text{aug}} \). MDS was originally proposed in psychometrics \[24\] for data visualization. Many adaptations of the method have been proposed for sensor localization. We use the so-called “s-stress” criterion \[25\]. Given an observed noisy matrix \( \mathbf{D}_{\text{aug}} \), s-stress(\( \mathbf{D}_{\text{aug}} \)) is the value of the following optimization program,

\[
\min \sum_{i,j} \left( \mathbf{D}_{\text{aug}}[i,j] - \mathbf{D}_{\text{aug}}[i,j]^\star \right)^2 \quad \text{s.t.} \quad \mathbf{D}_{\text{aug}} \in \text{EDM}^3.
\]

By \( \text{EDM}^3 \) we denote the set of EDMs generated by point sets in \( \mathbb{R}^3 \). We say that s-stress(\( \mathbf{D}_{\text{aug}} \)) is the score of the matrix \( \mathbf{D}_{\text{aug}} \), and use it to assess the likelihood that a combination of echoes...
The algorithm is summarized as

i) For every $d_{i_1,...,i_d}$, score($d_{i_1,...,i_d}$) ← s-stress($D_{2mh}$);

ii) sort the scores collected in score;

iii) compute the image-source locations;

iv) remove image sources that do not correspond to walls (higher-order by using step iii, ghost sources by heuristics); and

v) reconstruct the room.

Step iv is described in more detail in the SI Text. It is not necessary to test all echo combinations. An echo from a fixed wall will arrive at all of the microphones within the time given by the largest intermicrophone distance. Therefore, it suffices to combine echoes within a temporal window corresponding to the array diameter. This substantially reduces the running time of the algorithm. As a consequence, we can be less conservative in the peak-picking stage. A discussion of the influence of errors in the image-source estimates on the estimated plane parameters is provided in (15).

**Experiments**

We ran the experiments in two distinctly different environments. One set was conducted in a lecture room at EPFL, where our modeling assumptions are approximately satisfied. Another experiment was conducted in a portal of the Lausanne cathedral. The portal is nonconvex, with numerous nonplanar reflecting objects. It essentially violates the modeling assumptions, and the object was to see whether the algorithm still gives useful information. In all experiments, microphones were arranged in an arbitrary geometry, and we measured the distances between the microphones approximately with a tape measure. We did not use any specialized equipment or microphone arrays. Nevertheless, the obtained results are remarkably accurate and robust.

The lecture room is depicted in Fig. 6A. Two walls are glass windows, and two are gypsum-board partitions. The room is equipped with a perforated metal-plate ceiling suspended below a concrete ceiling. To make the geometry of the room more interesting, we replaced one wall by a wall made of tables. Results are shown for two positions of the table wall and two different source types. We used an off-the-shelf directional loudspeaker, an omnidirectional loudspeaker, and five nonmatched omnidirectional microphones. RIRs were estimated by the sine sweep technique (26). In the first experiment, we used an omnidirectional loudspeaker to excite the room, and the algorithm reconstructed all six walls correctly, as shown in Fig. 6B. Note that the floor and the ceiling are estimated near perfectly. In the second experiment, we used a directional loudspeaker. As the power radiated to the rear by this loudspeaker is small, we placed it against the north wall, thus avoiding the need to reconstruct it. Surprisingly, even though the loudspeaker is directional, the proposed algorithm reconstructs all of the remaining walls accurately, including the floor and the ceiling.

Fig. 6D and F shows a panorama view and the floor plan of the portal of the Lausanne cathedral. The central part is a pit reached by two stairs. The side and back walls are closed by glass windows, with their lower parts in concrete. In front of each side wall, there are two columns, and the walls are joined by column rows indicated in the figure. The ceiling is a dome ∼9 m high. We used a directional loudspeaker placed at the point L in Fig. 6F. Microphones were placed around the center of the portal. Also, in this case we do not have a way to remove unwanted image sources, as the portal is poorly approximated by a convex polyhedron. The glass front, numeral 1 in Fig. 6F, and the floor beneath the microphone array can be considered flat surfaces. For all of the other boundaries of the room, this assumption does not hold. The arched roof cannot be represented by a single height estimate. The side windows, numerals 2 and 3 in Fig. 6F, with pillars in front of them and erratic structural elements at the height of the microphones, the rear wall, and the angled corners with large pillars and large statues, all present irregular surfaces creating diffuse reflections. Despite the complex room structure with obstacles in front of the walls and numerous small objects resulting in many small-amplitude, temporally spread echoes, the proposed algorithm correctly groups the echoes corresponding to the three glass walls and the floor. This certifies the robustness of the method. More details about the experiments are given in the SI Text.

**Discussion**

We presented an algorithm for reconstructing the 3D geometry of a convex polyhedral room from a few acoustic measurements. It requires a single sound emission and uses a minimal number of microphones. The proposed algorithm has essentially no constraints on the microphone setup. Thus, we can arbitrarily reposition the microphones, as long as we know their pairwise distances (in our experiments we did not “design” the geometry of the microphone

![Fig. 6](image-url)
Further, we proved that the first-order echoes collected by a few microphones indeed describe a room uniquely. Taking the image source point of view enabled us to derive clean criteria for echo sorting.

Our algorithm opens the way for different applications in virtual reality, auralization, architectural acoustics, and audio forensics. For example, we can use it to design acoustic spaces with desired characteristics or to change the auditory perception of existing spaces. The proposed echo-sorting solution is useful beyond hearing rooms. Examples are omnidirectional radar, multiple-input–multiple-output channel estimation, and indoor localization to name a few. As an extension of our method, a person walking around the room and talking into a cellphone could enable us to both hear the room and find the person’s location. Future research will aim at exploring these various applications.

Results presented in this article are reproducible. The code for echo sorting is available at http://rr.epfl.ch.

ACKNOWLEDGMENTS. This work was supported by a European Research Council Advanced Grant–Support for Frontier Research–Sparse Sampling: Theory, Algorithms and Applications (SPARSAM) 247006. A.W.’s research is funded by the Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany.

Supporting Information

Dokmanić et al. 10.1073/pnas.1221464110

SI Text

1. Proof of the Theorem

**Theorem 1.** Consider a room with a loudspeaker and \( M \geq 4 \) microphones placed uniformly at random inside the feasible region. Then the set of first-order echoes uniquely specifies the room with probability 1. In other words, almost surely exactly one assignment of first-order echoes to walls describes a room.

**Proof:** It is sufficient to prove the claim for \( M = 4 \). Cases when \( M > 4 \) follow by considering any subset of four microphones. Draw independently and uniformly at random microphone locations \( \mathbf{r}_1, \ldots, \mathbf{r}_4 \) in the feasible region. To this particular choice of microphone locations we correspond vectors \( \mathbf{y}_k \) and \( \mathbf{y}_k \) as follows,

\[
y_{k,m} = ||\mathbf{s}_k - \mathbf{r}_m||^2 = ||\mathbf{s}_k||^2 - 2 \mathbf{s}_k \cdot \mathbf{r}_m + ||\mathbf{r}_m||^2,
\]

where \( \mathbf{s}_k \) is the location of the image source with respect to wall \( k \). We have in vector form

\[
\begin{bmatrix}
\hat{y}_{k,1} \\
\hat{y}_{k,2} \\
\hat{y}_{k,3} \\
\hat{y}_{k,4}
\end{bmatrix} = \begin{bmatrix}
\mathbf{r}_1 - \mathbf{s}_k \\
\mathbf{r}_2 - \mathbf{s}_k \\
\mathbf{r}_3 - \mathbf{s}_k \\
\mathbf{r}_4 - \mathbf{s}_k
\end{bmatrix} \begin{bmatrix}
||\mathbf{s}_k||^2
\end{bmatrix}, \quad \text{or} \quad \hat{y}_k = \mathbf{R} \mathbf{u}_k.
\]

Thanks to the condition \( \sum_{m=1}^4 \mathbf{r}_m = 0 \), we have that

\[
1^\top \hat{y}_k = -\frac{M}{2} ||\mathbf{s}_k||^2 \quad \text{or} \quad ||\mathbf{s}_k||^2 = -\frac{2}{M} \sum_{m=1}^M \hat{y}_{k,m}.
\]

The image source is found as

\[
\hat{\mathbf{s}}_k = \mathbf{S} \hat{\mathbf{y}}_k,
\]

where \( \mathbf{S} \) is a matrix satisfying

\[
\mathbf{S} \mathbf{R} = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0
\end{bmatrix}.
\]

It follows that

\[
\frac{2}{3} 1^\top \hat{y}_k + ||\mathbf{S} \hat{\mathbf{s}}_k||^2 = 0.
\]

Vector \( \hat{\mathbf{y}}_k \) corresponds to the \( k \)th wall, or \( k \)th image source (it is the correct permutation). We now show that wrong permutations cannot satisfy Eq. S7. We do it by replacing one, two, or three entries in \( \hat{\mathbf{y}}_k \) by wrong values and arguing that these are not good combinations. We choose

\[
\mathbf{S} = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0
\end{bmatrix} \mathbf{R}^\dagger,
\]

where \( \mathbf{R}^\dagger \) is the Moore–Penrose pseudoinverse of \( \mathbf{R} \). With this choice, any column submatrix of \( \mathbf{S} \) with \( n \leq 3 \) columns is rank \( n \) with probability 1.

1. (1 replacement). Without loss of generality, let us replace the fourth entry of \( \hat{\mathbf{y}}_k(\hat{\mathbf{y}}_k) \), by \( \hat{\mathbf{y}}_{k,k'} k' \neq k, \) and plug it into Eq. S7.

We can rewrite the equation as

\[
\alpha + \beta \hat{y}_{k',4} + \gamma \hat{y}_{k,4}^2 = 0,
\]

where \( \alpha, \beta, \) and \( \gamma \) do not functionally depend on \( \hat{\mathbf{y}}_{k',4} \), and \( \gamma \neq 0 \) with probability 1. For any realization of \( \hat{\mathbf{y}}_{k',1}, \ldots, \hat{\mathbf{y}}_{k',3} \), the distribution of \( \hat{\mathbf{y}}_{k',4} \) is continuous, thus the probability that it assumes any given value is zero (note that this is not true for \( \hat{\mathbf{y}}_{k',4} \)—for echoes coming from the same wall, knowing three of them constrains the fourth to two possible values). Therefore, the probability that \( \hat{\mathbf{y}}_{k',4} \) is one of at most two real roots of Eq. S9 is zero.

2. (2 replacements). Now we replace \( \hat{\mathbf{y}}_{k,3} \) and \( \hat{\mathbf{y}}_{k,4} \) by \( \hat{\mathbf{y}}_{k',3} \) and \( \hat{\mathbf{y}}_{k',4} \). We can have either (i) \( k' = k \) or (ii) \( k' \neq k \). We rewrite Eq. S7 as

\[
\begin{bmatrix}
\hat{\mathbf{y}}_{k,3} \\
\hat{\mathbf{y}}_{k',4}
\end{bmatrix} A \begin{bmatrix}
\hat{\mathbf{y}}_{k,3} \\
\hat{\mathbf{y}}_{k',4}
\end{bmatrix} + a^\top \begin{bmatrix}
\hat{\mathbf{y}}_{k,3} \\
\hat{\mathbf{y}}_{k',4}
\end{bmatrix} + a = 0,
\]

where \( A = [S; 3 : 4][S; 3 : 4] \) (with Matlab notation) is full rank with probability 1 and is positive semidefinite. Also, \( a \) and \( a \) do not functionally depend on \( \hat{\mathbf{y}}_{k,3}, \hat{\mathbf{y}}_{k',4} \). Locus of the roots of Eq. S10 is an ellipse. However, for any realization of \( \hat{\mathbf{y}}_{k,1} \) and \( \hat{\mathbf{y}}_{k,2} \), the distribution of \( \hat{\mathbf{y}}_{k,3}, \hat{\mathbf{y}}_{k,4} \) is continuous over some 2D subset of \( \mathbb{R}^2 \) both in cases i and ii. Therefore, the probability that it takes a value on the root ellipse of Eq. S10 is zero.

3. (3 replacements). Here we replace \( \hat{\mathbf{y}}_{k,2}, \hat{\mathbf{y}}_{k,3}, \hat{\mathbf{y}}_{k,4} \) with \( \hat{\mathbf{y}}_{k',2}, \hat{\mathbf{y}}_{k',3}, \hat{\mathbf{y}}_{k',4} \). If \( k' = k'' \), then the argument is the same as in the case of one replacement. If \( k' = k'' \) or \( k' = k'' \), but not all three are equal, then we can just repeat the argument for the case of 2 replacements (ii). Finally if they are all different, we write

\[
\begin{bmatrix}
\hat{\mathbf{y}}_{k,2} \\
\hat{\mathbf{y}}_{k,3} \\
\hat{\mathbf{y}}_{k,4}
\end{bmatrix} B \begin{bmatrix}
\hat{\mathbf{y}}_{k,2} \\
\hat{\mathbf{y}}_{k,3} \\
\hat{\mathbf{y}}_{k,4}
\end{bmatrix} + b^\top \begin{bmatrix}
\hat{\mathbf{y}}_{k,2} \\
\hat{\mathbf{y}}_{k,3} \\
\hat{\mathbf{y}}_{k,4}
\end{bmatrix} + b = 0.
\]

Again \( B = [S; 2 : 4][S; 2 : 4] \) is full rank with probability 1, so the locus of the roots of Eq. S11 is an ellipsoid. The set of values that \( \hat{\mathbf{y}}_{k,2}, \hat{\mathbf{y}}_{k,3}, \hat{\mathbf{y}}_{k,4} \) takes is again some 3D region in \( \mathbb{R}^3 \) and the probability that the triplet takes value on an ellipsoid is zero.

In conclusion, almost surely only one (correct) combination of echoes satisfies Eq. S7, so almost surely only one room corresponds to collected first-order echoes.

2. Experimental Setup

2.1. Equipment. We used a Lange D12A dodecahedron omnidirectional loudspeaker (Fig. S1A) and a two-way directional active monitoring loudspeaker Genelec 8030A (Fig. S1B). The
horizontal beam pattern of Genelec 8030A is depicted in Fig. S2. The horizontal directivity sonogram of Lange D12A is shown in Fig. S3.

To record the responses, we used five nonmatched Behringer ECM 8000 omnidirectional measurement microphones (Fig. S1C). The microphones and the loudspeaker were interfaced with a PC through a Motu 896HD unit (Fig. S1D) operating at a sampling frequency of 96 kHz.

2.2. Microphone Arrays. Table S1 contains the distances between the microphones in the experiments. Distances were measured between the tips of the omnidirectional microphones using a tape measure.

2.3. Measurement Technique. We measured the room impulse responses by the swept-sine technique (1). An excitation signal is played back over the chosen loudspeaker while simultaneously recording the signals arriving at the microphones. The played signal is a sine sweep with an instantaneous frequency varying exponentially with time,

\[ x(t) = \sin \left( \frac{\omega_1 T}{\ln(\omega_2/\omega_1)} \left( e^{\frac{t \ln(\omega_2/\omega_1)}{T}} - 1 \right) \right). \]  

[S12]

where \( \omega_1 \) is the start frequency, \( \omega_2 \) is the end frequency, and \( T \) is the total duration of the sweep in seconds. The recorded signals \( y(t) \) can be written in Fourier domain as \( Y(\omega) = H(\omega)S(\omega) \). Hence, the room transfer function \( (H(\omega)) \) can be computed by spectral division,

\[ H(\omega) = \frac{Y(\omega)}{S(\omega)}. \]  

[S13]

Inverse Fourier transforming \( H(\omega) \) yields the impulse response.

2.4. Remark About Loudspeakers. Peak-picking and RIR measurement techniques are out of the scope of this paper. Nevertheless, the loudspeaker size, build, and impulse response affect the quality of the estimation. This effect is indirect through the peak shape. Size is relevant, as we assume a point source. This assumption is satisfied for the directional speaker that has a well-defined acoustic center. However, the omnidirectional loudspeaker has widely placed drivers, so it is a poor approximation of a point source. We can partially compensate for the speaker size by assuming it is spherical and using the Huygens principle, but the structure of the impulse response still reflects the distributed drivers.

2.5. Delay of the Processing Chain. We measured the total delay of the processing chain to be 365 samples at the sampling frequency of 96 kHz. This offset must be accounted for when processing the impulse responses. In the case of the Lange omnidirectional loudspeaker, we used a smaller offset of 338 samples. The reason for this is to compensate for the loudspeaker size: At time 0, the sound wave is already at a distance \( R \) from the center of the loudspeaker, where \( R \) is the speaker radius. Therefore, the delay until the sound reaches some point in space is smaller than for the point source. Of course, this would be correct if the loudspeaker was a perfect sphere. In practice, we can only compensate the radius “in the mean.” That is also why in general, the results obtained with the omnidirectional loudspeaker are slightly less accurate than with the directional one.

2.6. Experiment in the Lausanne Cathedral. The measurements in the side portal of the Lausanne cathedral were challenging as a large part of the boundary surfaces are not flat (as assumed by the algorithm, and as has been the case in the classroom measurements).

The glass front (numeral 1 in Fig. 6F) and the floor beneath the microphone array can be considered flat surfaces. For all of the other boundaries of the room, this assumption does not hold. The arched roof cannot be represented by a single height estimate. The side windows (numerals 2 and 3 in Fig. 6F) with pillars placed in front and erratic structural elements at the height of the microphones, the rear wall, and the angled corners with large pillars and large statues, all present irregular surfaces creating diffuse reflections. Fig. 4 shows the details of the sidewall structure and the microphone arrangement. The waveform of a reflection from such diffusive architectural surfaces exhibits distinct differences compared with one from a large, flat surface. In general, such a reflection response is temporally spread and has a lower peak amplitude than an impulse containing the same energy (2). These characteristics are unfavorable for our algorithm because it is harder to detect peaks that actually belong to walls. Many of the detected peaks stem from reflections off small structural elements. The purpose of this experiment was to get an idea about the robustness of the echo-sorting algorithm to inputs from measurements made in environments that violate the assumptions made by the proposed algorithm.

The measurement procedure and equipment was the same as in the EPFL classroom measurements. As has been the case before, the microphones were not calibrated; the single channel preamp potentiometers had approximately equal settings.

To measure an impulse response of a room, a high-level broadband excitation signal is needed. Using an impulse as excitation signal, the recorded signal is immediately the impulse response. Unfortunately, impulsive sources (e.g., popping a balloon or firing a starter pistol) have poor repeatability, produce unpredictable spectra, and do not guarantee omnidirectionality (3). Assuming that the side portal is a linear and time invariant system, the required energy can be spread over time. We excite the room with a deterministic signal, and the room impulse response can be calculated from the signal recorded in the room. We applied the swept-sine technique as described in SI Text, section 2.3.

The loudspeaker used in this experiment is not omnidirectional. Therefore, in addition to the position of the loudspeaker and the microphones, the measured impulse responses depend on the orientation of the loudspeaker. This has been considered by placing the directional loudspeaker close to one wall. The microphone array was positioned in the lowered center part of the portal, which was surrounded on all four sides by stairs. We recorded the impulse responses with a randomly setup microphone arrangement.

The described effects of the architectural structures in the cathedral portal become apparent in the recorded impulse responses. Fig. S5 shows a comparison of impulse responses recorded in the lecture room and in the cathedral. We can see that the number of distinct peaks in the cathedral impulse responses is smaller than in the classroom measurement, and that the peaks in the cathedral RIR have lower amplitudes compared with the direct sound than the peaks in the classroom (the floorplan dimensions are comparable between the two cases, and the timescale was chosen accordingly).

3. Distances in Fig. 5

For aesthetic reasons the distances in Fig. 5 of the manuscript were specified to a single decimal place. Assuming the left lower corner of the room as origin, the exact microphone and loudspeaker positions are as follows.

The upper wall is at a distance 200/15 from the origin. Higher precision entries for the distance matrices are as follows,
4. Multidimensional Scaling

As pointed out, in the presence of noise it is not favorable to use the rank test on \( \mathbf{D}_{\text{aug}} \). A very good way (as verified through simulations) to deal with this nuisance is to measure how close \( \mathbf{D}_{\text{aug}} \) is to a true EDM. To measure the distance, we use multidimensional scaling (MDS) to construct a point set in a given dimension (either 2D or 3D), which produces the EDM “closest” to \( \mathbf{D}_{\text{aug}} \).

MDS was originally proposed in psychometrics as a method for data visualization (4). Many variations have been proposed to adapt the method for sensor localization. We use the s-stress criterion as proposed by Takane et al. (5). Given an observed noisy matrix \( \mathbf{D} \), the s-stress criterion is

\[
s(\mathbf{D}) = \text{minimize} \sum_{i,j} \left( d_{ij}^2 - d_{ij}^2 \right)^2
\]

subject to \( \mathbf{D} \in \text{EDM}^2 \).

We call \( s(\mathbf{D}) \) the score of matrix \( \mathbf{D} \). By EDM\(^2 \) we denote the set of EDMs with embedding dimension 2 (produced by point sets in 2D). In the 3D case, EDM\(^2 \) is replaced by EDM\(^3 \).

From now on, we assume that the target space is \( \mathbb{R}^2 \). The 3D adaptation is immediate. If we associate to each point in \( \mathbb{R}^3 \) a coordinate vector \( \mathbf{x} = (x_i, y_i, z_i)^T \), we have that \( d_{ij}^2 = \|\mathbf{x}_i - \mathbf{x}_j\|^2 = (x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2 \). Thus, the s-stress criterion can be rephrased as

\[
s(\mathbf{D}) = \text{minimize} \sum_{i,j} \left[ (x_i - x_j)^2 + (y_i - y_j)^2 - d_{ij}^2 \right]^2.
\]

The objective function in Eq. S14 is not convex. However, it has been shown to have less local minima compared with other MDS criteria (5). Furthermore, it yields a meaningful definition of the distance of a matrix from an optimal EDM.

To further skip the local minima of Eq. S14, we use coordinate alternation for finding the optimal EDM : we compute Eq. S14, by first minimizing over \( x_i \) and then over \( y_i \). Although this approach is suboptimal compared with simultaneous minimization with respect to \( x, y \), it leads to simpler computations.

Assuming that \( x_i \) has to be updated by \( \Delta x_i \) to give the minimum of \( s(\mathbf{D}) \), we will have

\[
s(\mathbf{D})^{(k+1)} = \sum_{i=1}^{n} \left[ (x_i^{(k)} + \Delta x_i^{(k+1)} - x_j^{(k)})^2 + (y_i^{(k)} - y_j^{(k)})^2 - d_{ij}^2 \right]^2.
\]

where \( \Delta x_i^{(k)} \) returns the value at iteration \( k \). Taking the derivative of \( s(\mathbf{D})^{(k+1)} \) with respect to \( \Delta x_i^{(k+1)} \), we will have

\[
\frac{\partial s(\mathbf{D})^{(k+1)}}{\partial \Delta x_i^{(k+1)}} = 4n \left( \Delta x_i^{(k+1)} \right)^3 + 3 \sum_{j=1}^{n} \left( x_i^{(k)} - x_j^{(k)} \right) \left( \Delta x_i^{(k+1)} \right)^2
+ \sum_{j=1}^{n} \left[ 3 \left( x_i^{(k)} - x_j^{(k)} \right) \left( y_i^{(k)} - y_j^{(k)} \right) - d_{ij}^2 \right] \Delta x_i^{(k+1)}
+ \sum_{j=1}^{n} \left[ \left( x_i^{(k)} - x_j^{(k)} \right)^3 + \left( x_i^{(k)} - x_j^{(k)} \right) \left( y_i^{(k)} - y_j^{(k)} \right) \right] \Delta x_i^{(k+1)}
- \left( x_i^{(k)} - x_j^{(k)} \right) \Delta d_{ij}^2.
\]

Setting Eq. S16 to zero yields at most three real solutions, and comparing the value of \( s(\mathbf{D})^{(k+1)} \) for the results gives the optimal value for \( \Delta x_i^{(k+1)} \).

The complete optimization procedure is summarized in Algorithm S1.

**Algorithm S1. Coordinate alternation for s-stress optimization**

**Input:** Symmetric and zero-diagonal matrix \( \mathbf{D} \)

**Output:** Estimate positions: \( \mathbf{x} \) and \( s(\mathbf{D}) \)

1. Assume an initial configuration for the points \( \mathbf{x}^0 \)
2. repeat
3. for \( i = 1 \) to \( n \) do
4. Assume the configuration of the points different from \( i \) fixed,
5. Update \( x_i \) using the \( i \)th row of \( \mathbf{D} \),
6. Update \( y_i \) using the \( i \)th row of \( \mathbf{D} \),
7. end for
8. until convergence or maximum number of iterations is reached.

**Algorithm S2. Room reconstruction procedure**

**Input:** Candidate images \( \mathbf{s}_1, \ldots, \mathbf{s}_p \), loudspeaker location \( \mathbf{s}_0 \), distance threshold \( \epsilon \)

**Output:** Room vertices

1. \( \{\mathbf{s}_1, \ldots, \mathbf{s}_p\} \leftarrow \text{SortByDistanceFromLoudspeaker}(\{\mathbf{s}_1, \ldots, \mathbf{s}_p\}) \)
2. \( \text{deleted}(1 ; p) \leftarrow \text{false} \)
3. for \( i = 1 \) to \( p \) do
4. if \( \forall j, k < i, j \neq k \) s.t. \( \|\text{Combine}(\mathbf{s}_j, \mathbf{s}_k) - \mathbf{s}_i\| < \epsilon \) then
5. \( \text{deleted}(i) \leftarrow \text{true} \)
6. else if Plane(\( \mathbf{s}_i \)) intersects the current room then
7. Add Plane(\( \mathbf{s}_i \)) to the current set of planes,
8. else
9. \( \text{deleted}(i) \leftarrow \text{true} \)
10. end if
11. end for

5. Room Reconstruction Procedure

The echo-sorting algorithm outputs a list of image sources. Some of these image sources are first-order images that we use to reconstruct the room. Some of the output image sources are higher-order sources, and we need to detect them and remove them from the list. As explained in the text, higher-order image sources are obtained as certain “combinations” of lower-order ones—a fact that we use to discriminate between them, as explained below.

We process the candidate image sources in the order of increasing distance from the loudspeaker. If the current image...
source cannot be obtained as a combination of closer sources, we add the corresponding plane (halfspace) to the list of halfspaces whose intersection determines the final room.  

Beyond the “combining criterion,” if the halfspace (which is really an inequality) that we are adding does not change the room, we discard the corresponding image source. We also do it if the new inequality perturbs the room only slightly.

This procedure is summarized in Algorithm S2. The following definition is used in the algorithm ($s$ is the loudspeaker):

$$\text{Combine } (s_1, s_2) \triangleq s_1 + 2\langle p_2 - s_1, n_2 \rangle n_2, \quad [S17]$$

where $p_2 = (s + s_2)/2$ is a point on the (hypothetical) wall defined by $s_2$, that is, a point on the median plane between the loudspeaker and $s_2$. The outward pointing unit normal is defined as $n_2 = (s_2 - s)/|s_2 - s|$. Room is defined as the intersection of halfspaces generated by the first-order image sources. With the above notation, halfspace corresponding to the image source $s_i$ is defined by

$$\{x : \langle n_i, x \rangle \leq \langle n_i, p_i \rangle\}. \quad [S18]$$

The plane corresponding to the image source $s_i$ is denoted simply by plane($s_i$).

6. Negative Answer to Kac’s Question

The reader might be interested by the construction of the counterexample to Kac’s question. Here, we explain a counterexample presented by Gordon and Webb (6). The beauty of their example is that elementary means suffice to understand why the two geometrically distinct drums (the same as those shown in Fig. 1) have the same resonant frequencies. However, how to systematically arrive at this construction (or other isospectral drums) is far more involved and requires the knowledge of advanced group representation theory (see the references in the paper).

The homogeneous Helmholtz (time-harmonic wave) equation on a domain $D$ with clamped boundary is given as

$$\Delta \phi + \lambda \phi = 0 \quad [S19]$$

$$\phi(x) = 0, \quad x \in \partial D. \quad [S20]$$

The solution needs to satisfy both Eq. S19 and Eq. S20. On compact domains, this equation admits the solution only for countably many eigenvalues $\lambda$, and the set of all admissible $\lambda$’s is denoted as the spectrum. Two domains are called “isospectral” if their spectra coincide (counting multiplicities). We note that the actual frequency is proportional to $\sqrt{\lambda}$, not to $\lambda$.

To understand the counterexample we need to use two properties of the solutions to the above equation,

i) Linearity: Linear combination of solutions is again a solution; and

ii) Reflection principle: If we have a solution on a domain bounded by a straight line segment with the clamped (Dirichlet) boundary condition, we can extend the domain and the solution by mirroring it over the line segment and changing the sign. This procedure ensures that the solution continues smoothly into the mirrored domain.

Now consider the two drumheads in Fig. S6 (these are the same as in Fig. 1). The drums are segmented and annotated as in Gordon and Webb (6). Let the vibrations of $D1$ be described by a function $\phi$ supported on the drum. The function $\phi$ satisfies both Eq. S19 and Eq. S20, for a given $\lambda$. Also, as indicated in Fig. S6, let $A, B, \ldots, G$ denote the restrictions of $\phi$ to corresponding triangular segments.

There happens to be a way to “transplant” the waveform from $D1$ to $D2$, so that the resulting waveform on $D2$ still satisfies both the Helmholtz Eq. S19 and the boundary condition Eq. S20. This transplantation is effected by placing linear combinations of $A, B, \ldots, G$ on $D2$, as indicated in Fig. S6, while observing the edge colors to ensure proper orientations.

We can check that the transplanted waveform indeed satisfies Eq. S19 and Eq. S20. Consider for example triangles $A+C+E$ and $-A+D+F$ on $D2$. We require that the corresponding waveforms combine smoothly over the blue edge. Triangles $C$ and $D$ share the blue edge on $D1$. The same holds for triangles $E$ and $F$. This means that they combine smoothly on $D1$ so $C+E$ and $D+F$ will combine smoothly on $D2$ as well. Now observe that the blue edge of $A$ on $D1$ is the boundary edge, so $A$ vanishes along the blue edge.

By reflection principle we can continue $A$ smoothly over the blue edge by mirroring it and multiplying by $-1$. Finally, this implies that $A+C+E$ and $-A+D+F$ will stitch smoothly. To check that the boundary conditions are satisfied, consider for example the triangle $-A+B+G$ and its red boundary edge. Triangles $A$ and $B$ share the red edge in $D1$, so they necessarily have the same value on the red edge. Thus, $-A+B$ is zero over the edge. In triangle $G$, red edge is the boundary edge, so $G$ is zero on that edge, and $-A+B+G$ must be zero on the boundary edge. It is easy to check that all of the triangles in $D2$ satisfy Eq. S19 and Eq. S20.

We showed that the Eq. S19 holds with the same $\lambda$ for both drums. Therefore, every resonant mode of $D1$ is also a resonant mode of $D2$. As we can also do a reverse transplantation procedure, every resonant mode of $D2$ is a resonant mode of $D1$, thus the two sets coincide, and the drums are isospectral.

Fig. S1. Equipment used in the experiments: (A) Lange D12A omnidirectional loudspeaker. (B) Genelec 8030 directional loudspeaker. (C) Behringer ECM 8000 omnidirectional microphone. (D) Motu 896HD unit.

Fig. S2. The upper curve group shows the horizontal directivity characteristics of the 8030A measured at 1 m. The lower curve shows the system’s power response.

Fig. S3. Directivity sonogram of D12A polar response with 1° resolution. Measurements made at 1 m distance.
Fig. S4. Details of the cathedral side portal (Left), microphone arrangement in the cathedral measurements (Right).

Fig. S5. Comparison of room impulses responses recorded in the lecture room and in portal of the cathedral.

Fig. S6. Isospectral drums with annotations as in (6).
Table S1. Microphone distances in centimeters used in experiments 1, 2, and 3

<table>
<thead>
<tr>
<th></th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>00.0</td>
<td>57.0</td>
<td>43.0</td>
<td>51.6</td>
<td>21.0</td>
<td>00.0</td>
<td>19.0</td>
<td>32.5</td>
<td>29.0</td>
<td>24.5</td>
<td>00.0</td>
<td>31.0</td>
<td>42.0</td>
<td>52.5</td>
<td>50.0</td>
</tr>
<tr>
<td>M2</td>
<td>57.0</td>
<td>00.0</td>
<td>25.0</td>
<td>43.5</td>
<td>44.0</td>
<td>19.0</td>
<td>00.0</td>
<td>42.5</td>
<td>45.5</td>
<td>40.0</td>
<td>31.0</td>
<td>00.0</td>
<td>54.0</td>
<td>54.5</td>
<td>31.0</td>
</tr>
<tr>
<td>M3</td>
<td>43.0</td>
<td>25.0</td>
<td>00.0</td>
<td>29.3</td>
<td>25.0</td>
<td>32.5</td>
<td>42.5</td>
<td>00.0</td>
<td>49.0</td>
<td>25.0</td>
<td>42.0</td>
<td>54.0</td>
<td>00.0</td>
<td>51.5</td>
<td>49.0</td>
</tr>
<tr>
<td>M4</td>
<td>51.6</td>
<td>43.5</td>
<td>29.3</td>
<td>00.0</td>
<td>32.3</td>
<td>29.0</td>
<td>45.5</td>
<td>49.0</td>
<td>00.0</td>
<td>27.5</td>
<td>52.5</td>
<td>54.5</td>
<td>51.5</td>
<td>00.0</td>
<td>42.5</td>
</tr>
<tr>
<td>M5</td>
<td>21.0</td>
<td>44.0</td>
<td>25.0</td>
<td>32.3</td>
<td>00.0</td>
<td>24.5</td>
<td>40.0</td>
<td>25.0</td>
<td>27.5</td>
<td>00.0</td>
<td>50.0</td>
<td>31.0</td>
<td>49.0</td>
<td>42.5</td>
<td>00.0</td>
</tr>
</tbody>
</table>

Table S2. Exact position of microphones and the loudspeaker in Fig. 5

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mic1</td>
<td>175/15</td>
<td>85/15</td>
<td>150/15</td>
</tr>
<tr>
<td>Mic2</td>
<td>220/15</td>
<td>100/15</td>
<td>165/15</td>
</tr>
<tr>
<td>Mic3</td>
<td>190/15</td>
<td>100/15</td>
<td>150/15</td>
</tr>
<tr>
<td>Mic4</td>
<td>220/15</td>
<td>70/15</td>
<td>150/15</td>
</tr>
<tr>
<td>Mic5</td>
<td>190/15</td>
<td>70/15</td>
<td>150/15</td>
</tr>
<tr>
<td>LSPK</td>
<td>100/15</td>
<td>140/15</td>
<td>150/15</td>
</tr>
</tbody>
</table>