This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Puomio, Otto; Lokki, Tapio

Common Image Source Search for Multiple Spatial Room Impulse Response Measurements

Published in:
2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)

DOI:
[10.1109/I3DA48870.2021.9610847](https://doi.org/10.1109/I3DA48870.2021.9610847)

Published: 10/09/2021

Document Version
Peer reviewed version

Please cite the original version:
Puomio, O., & Lokki, T. (2021). Common Image Source Search for Multiple Spatial Room Impulse Response Measurements. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)* (pp. 1-10). [9610847] IEEE. <https://doi.org/10.1109/I3DA48870.2021.9610847>

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Common Image Source Search for Multiple Spatial Room Impulse Response Measurements

1st Otto Puomio

*Department of Computer Science and
Department of Signal Processing and Acoustics
Aalto University
Helsinki, Finland
ORCID 0000-0001-8749-2674*

2nd Tapio Lokki

*Department of Signal Processing and Acoustics
Aalto University
Helsinki, Finland
ORCID 0000-0001-7700-1448*

Abstract—Image source reversion algorithms estimate room geometry from measured spatial room impulse responses by locating image sources. However, most of the methods have been limited to a single loudspeaker position and to convex rooms. Earlier, we have proposed a method that combines image sources from multiple receiver locations to find more image sources accurately even in concave rooms. Here, we extend the method to cope with multiple sound sources, thus the image source search can utilize measurements from multiple source and receiver positions simultaneously. The search method is tested with two measurement datasets and is found to improve the result compared to the previous algorithm. For the future applications, we also propose methods for generating a concave room model from the found reflection planes, for estimating material filters for each surface, and for interpolating between measured source locations.

Index Terms—spatial room impulse responses, room acoustics, early reflections, image sources, image source reversion

I. INTRODUCTION

In recent years, virtual acoustics field has shown an increasing research interest on transparent room modeling. Transparent room modeling aims at reproducing a real-life acoustic space so authentically that the listener cannot distinguish a physical sound event from a virtual one. For this, it is especially important to correctly model early reflections (ER) that help the listener to sense the shape and size of the room. These sensations derive from ER timings and directions, which calls for resolving the room geometry. When the geometry is known, the ERs can be simulated for all positions inside the model. In other words, achieving this enables both listener movement in 6 degrees of freedom and placing virtual sound source (SS) freely in the space.

Room geometry can be estimated from measured room impulse responses (RIR) by image source (IS) reversion [1]–[6]. The idea is based on the image source method [7], [8] where a reflecting surface creates a mirror image of the original SS. This also works the other way; IS reversion resolves the wall positions by locating the ISs generated for a known SS position. The reflecting surface is then placed halfway between the two points. Generally, the IS reversion methods use measurement array geometry and ER times-of-arrival (ToA) to locate first-order ISs in the acoustic scene.

Some of the methods also introduce notable extensions to this procedure. Ribeiro et al. [2] used second- and third-order reflections to validate found wall positions, while Tervo and Tossavainen [1] located walls also from higher-order ISs. Remaggi et al. [4] were first to introduce direction-of-arrival (DoA) to IS reversion. By doing this, they could locate the receiver microphone array and image sources by only knowing the SS positions. They also introduced mean and median-based methods to estimate wall positions more accurately by using multiple loudspeakers. Finally, Crocco, Trucco and Del Bue [5] resolved the reflector positions without knowing the measurement array geometry. However, all the aforementioned methods assume a convex space in one way or another, which also limit the spaces they can analyze.

Our previous work [9] touched on the IS reversion by locating ISs from multiple measured spatial room impulse responses (SRIR; i.e., RIRs containing spatial information). Similar to [4], we used ToAs and DoAs to locate the receivers. However, there were also multiple differences. The ToAs were resolved from the distance between the measured SS and fitted receiver positions, which allowed us to apply the method on legacy concert hall measurements. The method also utilized multiple receivers to present the ISs as multivariate normal (MVN) distributions instead of point detections, thus allowing to present the findings with a visible certainty measure. Moreover, the method did not try to reconstruct the room geometry, but only aimed at detecting as many ISs as possible. The implementation was still bound to one SS at a time, losing a great amount of potential data to find the ISs.

This paper is a continuum to the earlier research, aiming at modeling the room as plane reflectors and combining all source-receiver measurements to a single room model. The first contribution is a novel method to combine SRIR measurements from multiple sources and receivers by translating the detected reflections to a common coordinate frame. This leads to the second contribution, which is the ability to map concave spaces as well. Lastly, the algorithm constructs an IS tree from the found ISs, effectively allowing SS translation and hinting which ERs are prominent enough for rendering. The presented method is validated with two case studies, one being a concave rectangular room and the other a concave

space.

The article is organized as follows. Section II introduces the reader to the IS translation, which is the basis for the actual method presented in section III. The case studies are presented in section IV and discussed in section V. Finally, we propose the directions for future work in section VI and conclude the article in section VII.

II. IMAGE SOURCE TRANSLATION

As mentioned before, IS reversion is based on the image source method. This method is a geometrical model that presents specular reflections as reflected images of either the SS or the receiver. The analogy is possible because the incoming and outgoing angles of the sound event are equal w.r.t. the surface in a specular reflection. Instead of resolving the reflection angle, one can reflect the SS (or the receiver) w.r.t. the reflecting surface, effectively presenting the reflection as a straight line instead. In simple terms, the receiver 'sees' the SS through a 'mirror' surface, the reflection being the IS. As IS reversion tries to solve the wall position by exploiting this property, it is essential to thoroughly understand the image source method.

IS location $\mathbf{s}_i^j$ can be determined for the i th sound source position $\mathbf{s}_i$ w.r.t. the j th wall as follows:

$$\mathbf{s}_i^j = \mathbf{s}_i - 2 \mathbf{n}_j (\mathbf{n}_j \cdot (\mathbf{s}_i - \mathbf{p}_j)) \quad (1)$$

where $\mathbf{p}_j$ and $\mathbf{n}_j$ are the position and normal vector of the j th wall plane, respectively. Similarly, the translation vector $\mathbf{t}_i$ can be mirrored w.r.t. the j th wall as follows:

$$\mathbf{t}_i^j = \mathbf{t}_i - 2 \mathbf{n}_j (\mathbf{n}_j \cdot \mathbf{t}_i) \quad (2)$$

where $\mathbf{t}_i^j$ is the reflected translation vector. The mirroring operation can also be presented in matrix form $\mathbf{M}_j$ as

$$\mathbf{M}_j = \mathbf{I} - 2 \mathbf{n}_j \mathbf{n}_j^T, \quad (3)$$

where $\mathbf{I}$ is an identity matrix. Matrix form makes chaining the mirroring operations of Equation (2) easy:

$$\mathbf{t}_i^{j_1, \dots, j_n} = \mathbf{M}_{j_n} \mathbf{M}_{j_{n-1}} \dots \mathbf{M}_{j_1} \mathbf{t}_i. \quad (4)$$

where $\mathbf{t}_i^{j_1, \dots, j_n}$ is a translation vector mirrored from walls $j_1, \dots, j_n$ (in order starting from j_1). The chaining also makes operating higher-order ISs efficient; their mirroring matrix can be calculated in advance and then used to translate the ISs w.r.t. the SS position they have originally been created for.

In IS reversion, one resolves the wall positions and normals from IS positions instead. The solution is to assume that the wall resides halfway between the SS and its IS. In other words, the wall position $\mathbf{p}_j$ and normal $\mathbf{n}_j$ are determined for a SS $\mathbf{s}_i$ and an IS $\mathbf{s}_i^j$ as follows

$$\begin{cases} \mathbf{p}_j &= \mathbf{s}_i + \frac{1}{2}(\mathbf{s}_i^j - \mathbf{s}_i) \\ \mathbf{n}_j &= (\mathbf{s}_i - \mathbf{s}_i^j) / \|\mathbf{s}_i - \mathbf{s}_i^j\|. \end{cases} \quad (5)$$

where $\|\cdot\|$ is a 2-norm operator.

Fig. 1. Common image source search flowchart. The algorithm presented in [9] provides the acoustic scene information. Common image source search then iteratively searches for common image sources until no new candidates are found. Finally, the common image source candidates form an image source tree that enables interpolating sound source positions properly.

III. METHODS

Fig. 1 illustrates a flowchart of the presented algorithm. The flowchart has been divided into three stages. The first stage analyzes the acoustic scene from SRIRs measured between multiple SSs and receivers. The acquired information is then passed to the second stage, which applies common image source search (CISS) algorithm. The CISS algorithm searches the data for ISs shared by multiple SRIR measurements. The found ISs form so-called Common Image Sources (CIS) that present the ISs in the CSS coordinate frame. Also like ISs, CISs are associated with reflecting walls that are further utilized to find higher-order CISs. In the third and final stage,

the found CISs are used to construct an IS tree. IS tree allows interpolating the found CIS positions to correspond to the different SS positions in the scene. The three stages are also divided into smaller steps, which are discussed further in the following sections.

A. Scene analysis

Scene analysis stage applies the analysis algorithm presented in [9]. The only difference is that the search algorithm finds ISs for each SS separately. As the other algorithms are identical, therefore the reader is instructed to consult [9] for more detailed method descriptions. Nonetheless, the algorithm is also described briefly below.

Scene analysis consists of two steps. First, one measures SRIRs for multiple SS and receiver positions in the acoustic scene. The SRIRs are then used to reconstruct the source-receiver array geometry, which sets the measurements to a common coordinate frame. The SRIRs are also analyzed for ER features that describe the properties of the incoming reflections. The found ER feature vectors are presented as ER objects, each containing feature descriptors for peak prominence and DoA stability. By setting thresholds for these two values, one can control the sensitivity and noise-robustness of the output ER objects. Finally, the scene analysis outputs calibrated SS and receiver positions as well as filtered ER objects to the CISS algorithm.

B. Common Image Source search

CISS aims at finding ISs from ER objects obtained from multiple source-receiver measurements. Combining the measurements gains two benefits over using data from a single SS. First, using multiple SSs grows the amount of data available for the search. Second, the SSs generate ISs only from the walls visible to their position, which is usually only a subset of all possible ISs in the scene. By combining multiple SS positions, one can combine these subsets and therefore construct a more complete description of the acoustic space. As seen in Equation (1), however, the task is not trivial as IS positions depend on the parent SS position. For this reason, one cannot combine the ER objects from different SSs directly. Instead, the objects need a separate preprocessing to be comparable with each other.

The CISS algorithm is divided into three steps. First, the ER objects from different SSs are translated to a common coordinate frame; then, the data is searched for CIS candidates; and if new candidates are found, the SS positions are mirrored based on the new CISs before repeating the process. The detailed descriptions of these three steps are as follows.

1) *Center sound sources*: The first step of the CISS algorithm transforms the ER objects of each SS to share a so-called common sound source (CSS) position. The centering has been visualized in Fig. 2. Before the operation, the SSs (colored triangles) have their own position s_i in the acoustic scene. The corresponding ISs (empty circles) are generated for the given wall (grey line) as specified in Equation (1). As the IS positions depend on the SS position, the ISs from different

Fig. 2. Three sound sources (colored triangles) generate image sources from a wall in the middle. The image sources are detected in the measurements and located as ER objects (empty colored circles). When the sound sources are transformed to a common source position (grey triangle), also the ER objects of the wall are transformed close to each other (filled circles).

SSs do not overlap even if they are generated for the same wall. On the other hand, if the SSs are translated to a single position (grey triangle), the ISs from the same wall should also share a common position after a corresponding transform (filled circles).

The centering is performed as follows. First, one selects the CSS position one wishes to move all the SSs. Here, this position is selected to be mean of all the SS reference positions measured at the scene. With this knowledge, one can determine the translations between each SS position and the CSS position. The translations are then applied to the ER objects using Equation (2), assuming a plane wall between each ER object and its corresponding SS. In case the assumed wall actually exists, ER objects from different SSs tend to converge towards a single position, effectively revealing a CIS. In case of noise or higher-order reflections, the transform does not converge the points similarly. This topic is further discussed in section III-B3.

2) *Find common image source candidates*: The second step finds CIS candidates from centered ER objects. The search is done in two parts. The first part groups the centered ER objects with Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [10]. To further connect the results to the real world, the second step inspects the found ER object clusters w.r.t. the reflection peaks in the corresponding SRIRs. These two steps aim at finding the matching ER peaks robustly and efficiently.

DBSCAN utilizes two parameters to adjust the grouping behavior: minimum number of neighboring points $N_{\min}$ and grouping distance ϵ . In short, $N_{\min}$ describes how many neighboring points must reside within ϵ to consider a point to form a cluster. However, selecting these parameters depends heavily on the dataset. Here for instance, the data quality is affected by the accuracy of the geometry calibration, the number of SRIR measurements and the centering of ER objects. Poor parameter adjustment was found to lead to either false detections or even failure to detect some or any true CISs. It is sometimes necessary to manually tune the parameters, but they can also be estimated for a good-quality dataset.

$N_{\min}$ is easier to estimate from the two parameters. In the

Fig. 3. Approximating the group distance from the centered data. First, one calculates an ascending k-dist graph (black curve). The graph is limited at $\epsilon_{\max}$ (here, $\epsilon_{\max} = 1$ m), followed by fitting a 7th order polynomial (yellow line) to the remaining data. The polynomial is then searched for the point where the positive curvature starts. The data around this point is used to fit a zero-curvature tangent (blue line). ϵ is selected as the first data point that deviates from that line by a predetermined amount ϵ_+ (here, $\epsilon_+ = 0.01$ m).

literature, a recommended limit is $N_{\min} = 2 \times \dim (= 6)$ [11]. However, this limit was found too strict for higher-order CISs. In the end, the best results were obtained by setting the limit higher for lower-order reflections and lower for higher-order ones.

Traditionally, ϵ has been selected for the whole dataset from a sorted k-dist graph [10]. A sorted k-dist graph is formed by measuring the distance between each point and its k'th nearest neighbor and sorting the distances to descending order. ϵ is then selected by searching for a 'valley' in this graph by hand. However, the manual approach does not suffice here as the centered ER object positions change between search rounds (see section III-B3). Instead, the ϵ is estimated automatically from the data.

Fig. 3 illustrates the automated ϵ selection from centered ER objects. The selection algorithm first generates an ascending k-distance graph (black line). The graph is then cut at user-defined maximum distance $\epsilon_{\max}$ and smoothed by fitting a 7th-order polynomial (yellow curve). The polynomial is then searched for a positive zero crossing in curvature. That point is used to determine a tangent of the graph, effectively creating a baseline for an increasing polynomial. Finally, the first data point deviating from the tangent by a pre-defined amount ϵ_+ defines the selected group distance ϵ .

The second part of the search algorithm filters the ER objects in each CIS by their ToA. Practically, each ER object corresponds to a single sound level peak in the measured SRIR. The position of this peak determines when the corresponding IS has been detected at the receiver array. In case the ER objects really belong to the same IS, their ToAs should also match in the measurements.

The ToA filtering is executed as follows. First, the reference source positions are reflected by the wall associated with the corresponding CIS. Then one calculates ideal ToAs between the mirrored reference positions and fitted receiver positions. The ToAs then point the CIS position in each contributing SRIR. Next, the ideal ToAs are matched with the ER objects that are temporally close to them in the SRIR ($\text{ToA} \pm 1$ ms in our case). A matching ER object has similar sound level and direction as the ER objects in the CIS. The found objects correspond to the ones stored in the CIS if the wall stored in the CIS is close to the correct one; in other cases, the set of matched ER objects is different. Finally, the old ER objects in the CIS are replaced by the new ones, the wall normal and reference position are recalculated, and the process is repeated until no changes occur.

3) *Higher-order common image sources*: As mentioned in section III-B1, Equation (2) does not converge ER objects associated with higher-order ISs. This is because the equation assumes a single wall between the source and its IS. In case of a higher-order reflection, there are two or more walls that affect the position of the ER object. In most cases, approximating these walls with a single wall does not lead to a correct result. Instead, one must unravel the walls one-by-one in order to properly locate the higher-order CISs.

The walls required for finding higher-order CISs can be approximated from already found CIS candidates. Similar to ER object translation, the wall is assumed to reside between the source and its IS, normal pointing towards the original source. Each wall then mirrors the reference source positions, CSS position and corresponding translations, followed by running the CISS again. The only difference to earlier is that the reference positions are now located outside the room, which affects the walls approximated between the reference positions and ER objects. Effectively, this corresponds to unraveling one wall from the IS path, making it possible to cluster second-order image sources. For higher-order sources, the reference positions are just reflected from two or more walls before grouping.

Running the CISS repeatedly may cause the algorithm to detect some of the CISs multiple times. These detections raise from three issues. First, the search is expected to find the parent CIS; second, special cases such as perpendicular or parallel walls can most probably be found in multiple searches; and third, higher-order reflections from parallel walls tend to be found already on earlier search rounds. To filter out the duplicates, the ER objects of CIS candidates are compared with each other. If the CISs share one or more objects, they are considered as duplicates. In that case, the CIS containing more ER points is spared.

C. Sound source interpolation

The CISs enable interpolation of SS positions in the acoustic scene. Basically, each CIS position represents an IS in the current CSS position. When the CSS is moved, CISs can be translated by reflecting the translation vector w.r.t. the corresponding walls using Equation 4. However, proper in-

Fig. 4. Reconstructing an image source tree. A sound source (grey triangle) generates two first-order and one second-order image sources (blue and yellow triangles). When reconstructing the tree, the wall of the yellow image source can be applied on blue image sources to discover their connection. This is possible because the distance to the wall is equal for both image sources.

terpolation also needs proper relations between the CISs in order to properly process higher-order reflections. In short, one needs to reconstruct an IS tree from the found CISs.

Actually, the CISS algorithm has already found loose relations between the CIS candidates. The walls of lower-order CISs have been used to determine the higher-order CISs, indicating the found higher-order CISs are related to the corresponding walls. However, the found relations are not necessarily right. For instance, a higher-order CIS might be found already in the first search round if it was generated by two parallel walls. As the proper relations are unknown, it is better to ignore the known relations and to build the IS tree from scratch

The IS tree is reconstructed as follows. As illustrated in Fig. 4, one selects a CIS candidate (yellow triangle) and mirrors all other CISs (blue triangles) w.r.t. the wall associated with it. The mirrored CIS candidates are then compared with unmirrored candidates. Knowing that the source and its IS always reside at an equal distance from the mirroring wall, overlapping candidates imply connection in the IS tree. The wall also determines the direction of the parent-sibling relationship. The parent IS is the CIS in front of the wall and the sibling IS is the one behind it.

Connecting the CIS candidates becomes more challenging when their positions are presented as multivariate normal (MVN) distributions. In that case, not only the approximated CIS position is reflected, but the covariance matrix of the candidate must be mirrored as well. In practice, the mirrored covariance matrix $[\Sigma^{j_1}]^{j_2}$ is resolved as follows:

$$[\Sigma^{j_1}]^{j_2} = \mathbf{M}_{j_2} \Sigma^{j_1} \mathbf{M}_{j_2}^{-1} \quad (6)$$

where Σ^{j_1} is the original covariance matrix of an IS from a wall j_1 and $\mathbf{M}_{j_2}$ is the mirror matrix of a wall j_2 . Basically, this equals to only mirroring the eigenvectors of Σ^{j_1} . In addition to mirroring Σ^{j_1} , one needs to compare two MVN CISs for similarity by calculating Matusita's measure ξ for 3-variate normal distributions [12]:

$$\xi = \frac{2^{3/2} |\Sigma^{j_1}|^{1/4} |\Sigma^{j_2}|^{1/4}}{|\Sigma^{j_1} + \Sigma^{j_2}|^{1/2}} \times \exp \left\{ -\frac{1}{4} (\mathbf{s}^{j_1} - \mathbf{s}^{j_2})^T (\Sigma^{j_1} + \Sigma^{j_2}) (\mathbf{s}^{j_1} - \mathbf{s}^{j_2}) \right\} \quad (7)$$

The above equation returns a value between 0 and 1 describing the similarity of the two CISs presented by MVNs $(\mathbf{s}^{j_1}, \Sigma^{j_1})$ and $(\mathbf{s}^{j_2}, \Sigma^{j_2})$, $\mathbf{s}^j$ being the CIS position and Σ^j its covariance matrix. The sensitivity of associating two CISs can therefore be controlled by selecting a similarity threshold; here, all CISs that evaluate $\xi > 0.05$ similarity were considered as a pair.

After resolving the IS tree, CIS translation becomes simple. Each CIS in the tree can precalculate the mirroring matrix $\mathbf{M} = \mathbf{M}_{j_n} \dots \mathbf{M}_{j_1}$ based on the walls associated with it. When the CSS is moved, one can first calculate the translation between the old and the new CSS position and then apply it on the CISs as suggested in Equation (4).

IV. EXPERIMENTS AND RESULTS

The presented algorithm was tested with two different datasets. The first dataset was an Immersive Sound Studio (ISS), a rectangular room having hard brick walls except for two surfaces. One of these surfaces has several windows and the other one is covered with thick absorption material; thus it is hardly reflecting any sound. The second dataset was a more irregularly shaped office coffee room. Both spaces had their (nonabsorbing) furniture moved to the sides of the room. Then the spaces were measured with five SS positions and 10 receiver positions in the ISS and 7 receiver positions in the coffee room. In the coffee room, the sound source, i.e., the loudspeaker was always turned towards the receiver. On the contrary, loudspeaker orientations were kept static in the ISS, thus the relative orientation was different for all receiver positions.

The obtained SRIRs were then analyzed by the algorithm. The input parameters are presented for both rooms in Table I. The upper part of the table describes the variables that are constant during the search. These variables are peak prominence and DoA stability thresholds applied during the scene analysis and $\epsilon_{\max}$ used when automatically determining ϵ for DBSCAN. The lower part of the table contains the variables that change between iterations. The first iteration runs with unmirrored reference SS positions, while the later iterations apply one or more walls to the ER object data. $N_{\min}$ and ϵ_+ both adjust the DBSCAN grouping, see section III-B2 for more details.

A. ISS

The found CISs for the ISS room are shown in Fig. 5a. The outline of the room has been drawn as a red wireframe

TABLE I
THE PARAMETERS USED BY THE COMMON IMAGE SOURCE SEARCH ALGORITHM FOR THE TWO TEST ROOMS.

	ISS		Coffee room	
Peak prominence (dB)	6		6	
DoA stability (samples)	10		7	
$\epsilon_{\max}$	2		1	
Iteration	$N_{\min}$	ϵ_+ (m)	$N_{\min}$	ϵ_+ (m)
1	11	0.001	7	0.01
2	9	0.005	4	0.025
3	5	0.015	4	0.05
4			4	0.075

Fig. 5. (Use U3D-enabled pdf reader for interactive content) Common image source search results for the ISS. (a) Common image sources (black MVN distributions) and walls (green diamonds, normal as a red line) found for the room; (b) image sources (colored MVN distributions) found by the common image source search; (c) image sources found by the search algorithm presented in [9] (min. 3 ER objects per image source). The image sources of different sound sources have been presented in different colors.

and the reference SS and fitted receiver positions are shown with triangles and rectangles, respectively. The found CISs are marked as black elliptical markers and the CSS position is pointed by the black crosshair. The algorithm has found the reflections in the $+x$ -y corner up to third order. The absorbing wall is also apparent as no reflections were found for the $-x$ wall.

The reconstructed IS tree for the ISS is presented in Table II, the CSS presented as CIS 1. The table shows the parent CIS, number of ER objects, applied $N_{\min}$, selected ϵ , and the

position of the CIS. The parent-child relationships are also visualized in the tree above the table. The found CIS relations appear to be reasonable, though the number of detected CISs is small. Also considering the room coordinates are axis-aligned, one can notice a small variation in the detected coordinates. For example, this offset is visible in the second-order reflection in the $-x+y$ corner of the room in Fig. 5a. The reflection is expected to locate itself on the CSS crosshair in the section and transverse projections¹, yet it is positioned below the marker.

The found CISs can also be presented as SS-specific ISs. This kind of presentation is illustrated in Fig. 5b, where ISs of each SS have been plotted in their unique color. For comparison, Fig. 5c visualizes the ISs found by the search algorithm presented in [9]. Since the search approach is different, the former can present ISs that have only one ER object (small colored crosshairs), while the latter has at least three ER objects in each shown IS. The presented algorithm appears to form a result cleaner than the one in comparison. On the other hand, the comparison algorithm seems to detect higher-order reflections in the $\pm y$ direction that goes unnoticed in the present one.

B. A coffee room

The found CISs for the coffee room are shown in Fig. 6. Here, the room has been visualized from four perspectives, but otherwise the markups are the same as in Fig. 5a. The number of found ISs is noticeably greater than in the ISS, although not all first-order ISs have been found. Nevertheless, only a small wall aligned in x direction is left completely unnoticed as four walls are detected from second-order ISs. There are also a couple of oddly placed CISs — the two higher-order CISs behind the $-x$ wall appear to position higher than the

¹These two projections can be found in the views pop-up menu of the interactive model.

TABLE II

IMAGE SOURCE TREE GENERATED BY COMMON IMAGE SOURCE SEARCH ALGORITHM FOR ISS. (N = NUMBER OF ER OBJECTS, $N_{\min}$ = NUMBER OF APPLIED ER OBJECTS, ϵ = SELECTED GROUP DISTANCE, AND X, Y, Z ARE THE CIS COORDINATES)

CIS	Parent	N	$N_{\min}$	ϵ (m)	X	Y	Z
1	—	—	—	—	1.9	1.0	1.2
2	1	45	11	0.352	1.9	1.1	-1.3
3	1	28	11	0.352	1.7	-3.0	1.3
4	1	23	11	0.352	1.9	1.0	5.8
5	1	13	11	0.352	2.3	5.6	1.4
6	2	21	9	0.452	1.7	-3.0	-1.3
7	3	17	9	0.452	1.9	-3.1	5.8
8	1	11	11	0.352	9.2	0.8	1.4
9	3	5	5	0.413	9.2	-3.5	0.5
10	4	11	9	0.461	9.1	0.9	6.1
11	7	6	5	0.376	9.1	-3.3	6.0

Fig. 6. Common image sources (black MVN distributions) and walls (green diamonds, normal as a red line) found for the coffee room.

associated first-order reflections. Also the third-order IS behind the $+x$ wall appears to behave in the same way.

The formed IS tree is presented in Table III. The algorithm appears to have found a good number of first and second order ISs in addition to five third-order ISs. However, the five CISs furthest away from the CSS are not located correctly in the tree. In other words, the CISs themselves are approximately correct, but their parent CIS could not generate such an IS from its position.

Finally, Figs. 7 and 8 show the ISs found by CISS and the comparison algorithm [9], respectively. Similar to the results, the presented algorithm provides cleaner results, but misses higher-order ISs when compared to the comparison algorithm. Additionally, the presented algorithm appears to find more ISs for the same reflection than the comparison one. This effect can be seen as reflected SS array pattern repeating outside the room boundaries.

V. DISCUSSION

The CISS appears to find the most prominent ISs reliably. These ISs are the ones generated from largest reflective surfaces in the room. This is expected as those surfaces are visible to the most sources and receivers. Also, the second-order CISs help in finding smaller surfaces. For instance, in the coffee room, the three $+y$ walls have been detected through second-order CISs. There can be two reasons for this. First, second-order reflections might have been captured by more receivers than the first-order reflections do from the same walls. Second, the walls might have been treated the way that the first-order reflection is more silent or scattered than the second-order one. This is probably true for the leftmost $+y$ wall in the coffee room; that wall is set up as a kitchen, making the surface more scattering than the larger walls around it.

TABLE III
IMAGE SOURCE TREE GENERATED BY COMMON IMAGE SOURCE SEARCH ALGORITHM FOR THE COFFEE ROOM. (N = NUMBER OF ER OBJECTS, $N_{\min}$ = NUMBER OF APPLIED ER OBJECTS, ϵ = SELECTED GROUP DISTANCE, AND X, Y, Z ARE THE CIS COORDINATES)

CIS	Parent	N	$N_{\min}$	ϵ (m)	X	Y	Z
1	—	—	—	—	4.2	4.5	1.2
2	1	32	7	0.420	4.2	4.5	-1.2
3	1	30	7	0.420	4.2	4.5	3.8
4	2	19	7	0.420	4.2	4.5	6.1
5	3	16	7	0.420	4.1	4.4	-3.9
6	4	11	7	0.420	4.2	4.4	-6.1
7	1	17	7	0.420	-4.0	4.6	1.2
8	2	13	4	0.264	-4.2	4.5	-1.0
9	1	8	7	0.420	4.0	-4.5	1.3
10	2	5	4	0.264	-3.3	12.4	-1.1
11	1	9	7	0.420	15.9	4.6	1.5
12	9	4	4	0.265	11.6	-4.7	1.2
13	7	6	4	0.260	-4.3	-4.1	0.8
14	11	4	4	0.254	15.7	11.2	1.0
15	7	6	4	0.260	9.7	17.5	1.0
16	13	5	4	0.310	20.2	-4.4	2.2
17 ^a	7	5	4	0.260	-15.6	4.4	2.5
18 ^a	11	10	4	0.260	24.1	4.1	1.6
19 ^a	4	4	4	0.260	-23.9	5.2	2.9
20 ^a	18	4	4	0.311	36.2	4.3	1.4
21 ^a	8	4	4	0.298	43.9	4.8	3.2

^aThe CIS has an incorrect parent

Fig. 7. Image sources found by the common image source search. The image sources of different sound sources have been presented in different colors.

Fig. 8. Image sources found by the search algorithm presented in [9] (min. 3 ER objects per image source). The image sources of different sound sources have been presented in different colors.

However, the second-order reflection from the floor only hits the kitchen cupboards that form a relatively flat surface.

The detection accuracy appears to deteriorate at higher distances. This is the most apparent in the coffee room where the five furthest CISs are located incorrectly in the IS tree. The false detections are probably caused by inaccurate wall parameter estimates, which in turn affect the translated ER object positions in later search rounds (see section III-B3). Even though a small error in the reflector position and normal is insignificant at short distances, the error becomes more significant the further the translated ER object is. Consequently, the distant CISs may then be incorrectly located, which prevents finding their parent reflections properly.

Highly absorptive and particularly small surfaces may be left unnoticed by the algorithm. The absorption case is apparent on the -x wall of the ISS; the ER peaks from the wall are strongly attenuated, thus its reflections are not detected.

The undetected surface case is in turn seen on the right -y wall of the coffee room. There, the small surface is located away from most of the SSs and receivers. Therefore, the wall only appears in a few measurements at best. In the end, the number of ER objects is too small to detect the CISs even from higher-order reflections.

Loudspeaker directivity may also have affected the receiver fit and the detected reflections. The measurement loudspeakers were highly directive and either statically positioned as in the ISS or always pointed at the receiver as in the coffee room. The DoAs are estimated the most accurately in the coffee room case where the direct sound peak is the most prominent. In the ISS case, however, the detected direction of the SS can vary due to the physical dimensions of the loudspeaker. This variance causes worse receiver fit than in the directed loudspeaker case, which in turn spreads the ER object clusters to a larger area. On the other hand, the loudspeaker

also emanates less high frequencies to the sides and behind the loudspeaker. Therefore, the ER peaks are not as prominent in those directions, which in turn reduce the probability of detecting the reflections in those directions. Both issues could be circumvented by reconstructing the SRIRs from multi-angle measurements as proposed in [13]. Receiver fit could also be easily improved by measuring the full audio chain delay by placing the measurement microphone very close to each measurement loudspeaker. This way, the receivers could be easily fitted by using the ToAs and DoAs of the direct sounds. The fit could be expected to improve, which would also improve the detection accuracy.

As expected, the CISS appears to add robustness to finding image sources when compared to the previous method presented in [9]. The found ISs appeared to have less covariance and conform to the SS array geometry better than the ones found with the comparison method, yet the new method also found less higher-order reflections. This phenomenon is probably because of the ToA filtering presented in section III-B2. The filtering practically limits the found ToAs within a certain spatiotemporal frame, which also limits the maximum covariance the CIS can have. While the presented procedure provides more accurate results, it also excludes less well-fitted ERs. As the ER position estimates get more inaccurate the further they are from the receiver, ToA filtering affects the furthest CIS candidates the most. On the contrary, the comparison procedure does not have a corresponding feature, but its grouping is solely limited by DBSCAN parameters. Consequently, the found ISs can be found later in time, yet they do not necessarily match the actual ERs in the scene. Therefore, a loss of few higher-order reflections does not appear as a bad trade-off for enhanced reliability.

Finally, the input parameters appeared to drastically affect the search performance. On one end, the algorithm was not able to find all the first-order reflections if either the $N_{\min}$ was set too high or ϵ_+ too low. On the other end, adding too much slack caused the search either to find nonexistent reflections or losing prominent ones due to two or more reflections being put to one group. Either way, selecting a grouping method other than DBSCAN could give more robust results.

VI. FUTURE WORK

Although the CISS finds reflecting planes from a generic space promisingly, there are still several steps to take before a complete six-degrees-of-freedom model. First, the planes must be converted to a room model in order to handle visibility properly; then one needs to resolve the material filters for each wall; and finally, the model and the filters need to be applied to generate arbitrary impulse responses within a room. Next, these three stages are shortly discussed to outline the work required to render the SRIRs with the presented analysis results.

A. Room model reconstruction

The current model presents the reflectors as infinite reflecting planes. However, as the planes are infinite, there is no

direct way to resolve occlusions from the other walls. This has been no problem before as the previous acoustic room reconstruction models have mostly assumed the space to be convex. Convex rooms have no obstructions by definition, so handling the visibility within the space has not been necessary. The visibility check is however required when moved to concave spaces, calling for a room model estimate to handle the visibilities properly.

One possible way to reconstruct the room model is to use found plane walls and known acoustic paths to generate the room. The walls are used to generate a set of corner point candidates by intersecting three planes with each other. Then the acoustic paths, namely direct sound and early reflection paths, are used to generate 'open space volume'. Basically, one would generate rays for each direct sound and ER and use them to generate a minimum volume around the rays. This volume would then be considered as an open space within the room, and the room itself would be a minimum set of corner points conforming to the rays.

B. Material filters

Material filters describe how the wall surfaces absorb the reflecting sound waves. A filter of a highly absorbing surface differs from a painted concrete wall, which absorbs very little sound. By modeling these differences, one might improve the quality of the rendered RIR.

The material could be analyzed from the measured SRIRs using the found CISs. Each CIS refers to multiple ER objects in different source-receiver measurements. The ER objects in turn locate the reflections in the RIRs. One can therefore form a specular reflection model from the ER objects located under each first-order CIS. The higher-order CISs are then combined from these modeled materials. In theory, combining different reflection angles from different measurements would even allow to approximate angle-dependent filters for the walls.

However, there are still several aspects that need to be solved for the materials. First of all, the detected absorption is influenced by the directivity of the SS. As a solution, one could either use an omnidirectional SS or simulate one by rotating a directive loudspeaker as in [1]. Secondly, the signal-to-noise ratio of the ERs degrades fast over time due to traveled distance, absorption, and diffuse reflections. This could be solved by averaging over multiple data points, yet this would call for more measured points. Thirdly, angle-dependent filter would also need more data than a convenient filter due to its complexity. Increasing the number of measurement positions manually would be inconvenient, thus exploring automated SRIR collection approaches such as [14] would be a reasonable next step.

C. Rendering

The model presented here follows a conventional rendering pipeline for spatial sound applications. The early reflection visibilities and delays would be first modeled using the SS and receiver positions and the resolved room geometry. The delays would then be used to read an input ring buffer for

the sound. To create a spatial output, first ERs could be directly spatialized by using head-related transfer functions and the ERs later in time would be combined to a higher-order Ambisonics signal. The late reverberation could be rendered with a feedback-delay network.

The new directions for the rendering would be how the ISs and material filters would be used. As the CIS positions w.r.t. the CSS position are known, it would be enough to transform the CISs to correspond to the virtual SS position. What comes to the directional material filters, one would need to interpolate reflections w.r.t. reflection angle. In that case, one could interpolate between two filters representing the two closest angles or, if the filter angles were dense enough, use the closest filter directly.

VII. CONCLUSIONS

This paper presents common image source search that locates image sources from multiple spatial room impulse response measurements, which are measured between multiple sound sources and receiver microphone arrays. The algorithm iteratively searches for first and higher order image sources by transforming the detected early reflection events to a common coordinate frame. By doing this, the search can locate image sources from multiple sound sources up to third order with a reasonable accuracy. The method also proved more robust in finding image sources than its predecessor, yet some of the higher-order image sources were lost in the process. Furthermore, the algorithm was able to reconstruct an image source tree from the detected image sources, although it failed to correctly position the ones that were far away from the common sound source. Nevertheless, the method proved that it is possible to combine measurements from multiple sound sources and receiver arrays in order to model even concave spaces. In the end, the authors suggested several future directions for this work, including concave room model reconstruction, directional material filters and a renderer for the presented acoustic model.

REFERENCES

- [1] S. Tervo and T. Tossavainen, "3D room geometry estimation from measured impulse responses," in *Proc. - ICASSP IEEE Int. Conf. Acoust. Speech Signal Process.* Kyoto, Japan: IEEE, August 2012, pp. 513–516.
- [2] F. Ribeiro, D. Florencio, D. Ba, and C. Zhang, "Geometrically constrained room modeling with compact microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1449–1460, 2012.
- [3] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *PNAS*, vol. 110, no. 30, pp. 12 186–12 191, 2013. [Online]. Available: <https://www.pnas.org/content/110/30/12186>
- [4] L. Remaggi, P. J. B. Jackson, P. Coleman, and W. Wang, "Acoustic reflector localization: Novel image source reversion and direct localization methods," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 2, pp. 296–309, 2017.
- [5] M. Crocco, A. Trucco, and A. Del Bue, "Uncalibrated 3D room geometry estimation from sound impulse responses," *J. Franklin Inst.*, vol. 354, no. 18, pp. 8678–8709, 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0016003217305458>
- [6] Y. Baba, A. Walther, and E. Habets, "3D room geometry inference based on room impulse response stacks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 5, pp. 857–872, 2018.

- [7] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.
- [8] J. Borish, "Extension of the image model to arbitrary polyhedra," *J. Acoust. Soc. Am.*, vol. 75, no. 6, pp. 1827–1836, 1984.
- [9] O. Puomio, N. Meyer-Kahlen, and T. Lokki, "Locating image sources from multiple spatial room impulse responses," *Appl. Sci.*, vol. 11, no. 6, 2021. [Online]. Available: <https://www.mdpi.com/2076-3417/11/6/2485>
- [10] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, E. Simoudis, J. Han, and U. Fayyad, Eds. American Association for Artificial Intelligence, 1996, pp. 226–231.
- [11] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu, "Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications," *Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 169–194, 06 1998, copyright - Kluwer Academic Publishers 1998; Last updated - 2020-11-18. [Online]. Available: <https://www.proquest.com/scholarly-journals/density-based-clustering-spatial-databases/docview/230142194/se-2?accountid=27468>
- [12] M. Minami and K. Shimizu, "Estimation of similarity measure for multivariate normal distributions," *Environ. Ecol. Stat.*, vol. 6, no. 3, pp. 229–248, Sep 1999. [Online]. Available: <https://doi.org/10.1023/A:1009678412953>
- [13] S. Tervo, J. Pätynen, and T. Lokki, "Acoustic reflection path tracing using a highly directional loudspeaker," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 245–248.
- [14] G. Götz, S. J. Schlecht, A. M. Ornelas, and V. Pulkki, "Autonomous robot twin system for room acoustic measurements," *J. Audio Eng. Soc.*, vol. 69, no. 4, pp. 261–272, 2021. [Online]. Available: <http://www.aes.org/e-lib/browse.cfm?elib=21033>