Can a Phone Hear the Shape of a Room?

Oliver Shih
Carnegie Mellon University
Pittsburgh, PA
oshih@alumni.cmu.edu

Anthony Rowe
Carnegie Mellon University
Pittsburgh, PA
agr@andrew.cmu.edu

ABSTRACT

Understanding the location of acoustically reflective surfaces in a room is a critical component in advanced sound processing. For example, intelligent speakers can use a room's acoustic geometry to improve playback quality, source separation accuracy, and speech recognition. In this paper, we present Synesthesia, a system for capturing the acoustic properties of a room using a single fixed speaker and a mobile phone that records audio at multiple locations. Using the arrival time of echoes, the system is able to reconstruct the position of reflective surfaces like walls and then estimate properties like surface absorption.

Previous work has shown how the acoustic room impulse response (RIR) of an environment can be used to analyze echoes within a space to reconstruct room geometry. The best current RIR-based approaches rely on high-end equipment and capturing an acoustic signal broadcast into space from a known fixed constellation of microphones. They also require the precise calibration and measurement of microphone positions. In addition, most approaches pose constraints on room geometries and limit the order of RIR to achieve accurate and consistent results. In this paper, we introduce a new approach that performs RIR imaging using a mobile phone that tracks its location with visual inertial odometry (VIO) to record a dense set of samples albeit with noise in their locations. We present a new approach that is able to relax several key assumptions on RIR and show through both experimentation and simulation that even with 20cm of uncertainty in the microphone locations provided by VIO, we are still able to reconstruct the room geometry with accurate shape and dimensions. We demonstrate this capability by prototyping a tool for acoustic engineers, that allows a user to view a room's estimated geometry and absorption overlaid on the actual sensed space with augmented reality.

CCS CONCEPTS

• **Theory of computation → Nonconvex optimization**; *Unsupervised learning and clustering*; • **Human-centered computing** → *Mobile computing*;

KEYWORDS

Active acoustic sensing, room reconstruction and mapping

IPSN '19, April 16–18, 2019, Montreal, QC, Canada

ACM ISBN 978-1-4503-6284-9/19/04…$15.00
https://doi.org/10.1145/3302506.3310407

ACM Reference Format:
Oliver Shih and Anthony Rowe. 2019. Can a Phone Hear the Shape of a Room?. In *The 18th International Conference on Information Processing in Sensor Networks (co-located with CPS-IoT Week 2019) (IPSN '19), April 16–18, 2019, Montreal, QC, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3302506.3310407

1 INTRODUCTION

The field of architectural acoustics is a branch in acoustic engineering that focuses on improving the sound quality within buildings. This could range from enhancing speech clarity in an auditorium to reducing background noise in a restaurant or simply improving the quality of music in a concert hall or recording studio. One of the main challenges in this field is understanding room impulse response (RIR) along with the location of various sound reflecting surfaces. This information can be exploited for a wide range of applications ranging from audio forensics [22] to creating 3D spatial sound effects [39]. The interaction between sound and the environment can be used by smart speakers [12, 15, 16], or in enterprise settings, to either improve music quality or tune beam-forming algorithms to enhance speech recognition [6, 13, 37]. In contrast to existing room mapping approaches like laser and depth sensors, acoustic sensing captures the surfaces that have the most significant impact on sound performance in space. For example, glass reflects sound but allows light to easily pass, and materials like felt absorbs sound but would be detected by vision or lasers.

Currently, when acoustic engineers optimize the sound properties of a space, they draw from a set of sound modification options like adding sound absorbers and structures to block noise, adjusting frequency levels, or leveraging electronic sound masking systems to combat various acoustic problems. For specialized listening areas like music halls, modeling tools can help optimize construction, but the fine-tuning of real-world sound performance typically requires an arduous trial and error process where installers evaluate various configurations of absorbers and reflectors either with measurement microphones at specific points in space or with a well-trained ear. The geometry of the space and the absorption coefficient of all surfaces plays a large role in a space's overall acoustics. This makes it extremely difficult to optimize acoustic properties, especially with a limited number of sampling points. In the case of smart speakers trying to sense the environment, they have the disadvantage of only being able to listen at a single point in space.

In this paper, we introduce Synesthesia[1], a system that takes the first steps towards providing acoustic engineers with the ability to accurately capture and visualize the reflection and absorption of sound within interior spaces through the use of a mobile phone as a receiver. With acoustic room geometry information (not just wall

[1] Synesthesia is named after the phenomena where one sense in a person triggers a reaction in another sensing system (e.g. seeing sound).

locations), smart speakers and high-end audio theater systems can better sense their environment to improve both their sound output quality as well as the ability to understand voice commands while playing music. They are currently limited with a single microphone that can only sense in a single fixed location. By using visual inertial odometry (VIO) on a smartphone (provided by platforms like ARKit [5]/ARCore [4]) we can precisely track a phone's relative location through space while simultaneously capturing a dense set of acoustic samples. Figure 1 shows an overview of Synesthesia that consists of a single fixed speaker array (i.e. a smart speaker) that generates a number of acoustic and ultrasonic chirps. The transmissions of chirps are synchronized with the mobile phone as the user walks around the space. Once a user has covered enough ground, our system can learn the RIR at each location to estimate the location of acoustic reflectors (like walls) based on echo arrival time and amplitude. After the acoustic room geometry has been reconstructed this model can be passed on as information to an audio processing system like a smart and adaptive speaker to improve sound quality. As an example, we use Synesthesia to create a heat map of acoustic absorption at a range of test frequencies projected on each surface in a room. Since the geometry is constructed relative to the VIO starting point of the phone, it is possible to overlay and visualize the final heat map using augmented reality. New versions of ARKit 2 support visual relocalization, so the phone can reload or share this information. This creates a powerful new way for users to explore the space seeing actual acoustic absorption mapped as colors in the environment. Though out of the scope of this paper, the same acoustic map may eventually be used to optimize speaker performance [29].

In the early 1900's, Sabine began to model the impact of people, frequency and the geometry of spaces on acoustics [32]. Significant follow-on work has explored modeling sound in indoor spaces to the point where it is possible to use the arrival time of echoes reflected off of walls to reconstruct room geometry [3, 8, 11, 17, 24, 30, 31, 36]. These approaches leverage the RIR to find the most likely position of walls in space using a set of speakers and microphones in a fixed and well-known configuration. For Synesthesia to achieve its goal of seamlessly allowing a phone to scan the space, we must relax a few of the key underlying assumptions from this body of previous work. First, our approach does not require prior knowledge of the number of reflective sources (typically walls) found in the room. Second, we do not assume that we receive all of the first echoes reflected off of surfaces. This is critical when dealing with mobile objects like people moving in a space. Third, we assume there are errors in the location estimates we received from our microphone placements provided by VIO. We transform the reconstruction problem into a multi-layered optimization problem using Euclidean distance matrix (EDM) properties, and apply techniques including semi-definite programming (SDP), combinatorial optimization, searching, and clustering to tackle the problem. Using a dense sampling of chirp recordings with relative positioning, we can create a high-resolution 3D image of reflective and absorbing surfaces in any given space. For the rest of the paper, the term "room geometry" refers to all of the acoustically reflective surfaces in a room.

In our prototype, the audio signals are transmitted from a Bluetooth triggered piezo speaker and recorded by a smartphone. The smartphone is initialized from the speaker's position and is used to trigger transmission and recording of test waveforms while annotating them with the visual odometry coordinates. Room geometry reconstruction and the absorption imaging are computed offline and then transmitted to an augmented reality phone application as a series of colored 3-Dimensional translucent polygons.

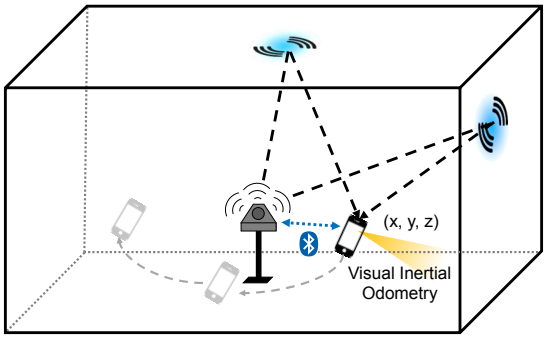


Figure 1: System overview

In summary, the main contributions of this paper are:

- An approach for acoustic room geometry reconstruction that assumes no prior knowledge about the number of surfaces and does not assume detection of all first-order echoes. Our approach is also robust to position error in recording data.
- We also present a platform that uses a single centrally located speaker module and visual inertial odometry on a mobile phone to simplify recording samples from a large number of microphone positions required to create an image of a space.
- Finally, we demonstrate the end-to-end system with an augmented reality tool that can visualize an estimate of the sound absorption coefficient of materials in a room.

2 RELATED WORK

In the 1960's, Kac famously lectured about a solution for solving the age-old physics problem of estimating the geometry of a drum based on the sound generated from striking the surface. More recently, this problem has been extended to estimating the shape of rooms based on their acoustic RIR [3, 8, 11, 17, 24, 30, 31, 36]. Acoustic geometry reconstruction typically assumes a set of microphones and speaker arrays with known locations. These transmit and receive pairs estimate the location of walls and obstacles based on the time delays of echoes. Most approaches rely on measuring the RIR and find the most likely location of walls based on the signals' time of arrival (TOA) [8, 11, 17, 30, 31, 36] or time difference of arrival (TDOA) [3, 24] of impulses, depending on whether the speakers and microphones are synchronized.

One approach models walls as planar surfaces tangent to the ellipsoid defined by the distance between transmitter/receiver pairs [3, 31]. To find the overlaps among multiple ellipsoids derived from noisy measurements, most techniques adopt Hough transform or RANSAC to reliably and efficiently find the best solution. However, they often require known microphone positions and the localization of the sound sources using the direction of arrival (DOA).

Another approach utilizes the Image Source (IS) model [1] to describe how the sound waves propagate to reduce the computation complexity of wall localization. In [24], the author exploits the

Figure 2: Experiment setup

constraints on convex polyhedral room geometry imposed by the combination of first-order and second-order reflections and presents a method to reconstruct room geometry from a single channel impulse response. Unfortunately, this method requires the detection of all first-order and second-order reflections, which is difficult in practice. In [11], Dokmanić et al. proposed using the properties of EDM to brute-force all combinations of echo arrival with multidimensional scaling (MDS) for estimating the location of image sources and their corresponding walls. To alleviate computational complexity, however, the system uses an array of 5 microphones with known positions to limit the number of echo combinations. In addition, the algorithm requires prior knowledge of the number of walls and the detection of all first-order echoes to eliminate higher-order echoes and correctly reconstruct room geometry. Similar follow-on work in [17] later transformed the echo combination problem into a maximal independent set listing problem in graphs that can be solved more efficiently using an exponential space algorithm. They used a rank-5 factorization method that was first proposed in [27] to directly compute the location of the transmitters and receivers in linear time. While promising in simulation, this approach requires at least ten sound sources and five microphones.

More recent work leverages a mobile device to replace multiple microphones [26, 38]. In [26], the author studied the possible room shapes that can be recovered using a mobile node, but assumed perfect localization and precise echo ranging. In [38], a commodity smartphone is used to achieve fine-grained reconstructions through short-range scanning. However, the method requires the user to walk a full loop closely to the internal room boundaries which is unsuitable for 3D reconstruction. To reliably measure the distance to walls, the smartphone also needs to be held in a specific position and follows a careful measurement gesture that is prone to error.

In this paper, we assume no prior knowledge of the number of reflective surfaces or the detection of all first-order echoes. We present a robust 3D reconstruction algorithm that utilizes semi-definite programming (SDP) to refine surfaces localization, a combinatorial optimization technique to cope with measurement uncertainty, and use a clustering algorithm exploiting geometry properties to deal with missing/spurious echoes. We also present a searching heuristic to reduce the overall computation complexity. Our system requires only one speaker and a commercial off-the-shelf smartphone that samples at multiple random locations in the room.

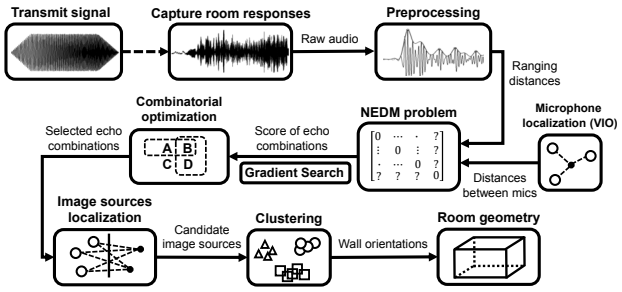


Figure 3: Overview of the room reconstruction algorithm

3 SYSTEM OVERVIEW

An overview of our experimental setup is shown in Figure 2. Our prototype consists of an omnidirectional tweeter speaker with custom hardware and a mobile device. The omnidirectional speaker design helps to maximize the number of reflections from each wall and thus reduces the total number of measurements required for reconstruction. Alternatively, one can rotate a directional speaker to reconstruct a subset of walls at a time, and later stitch the results together. Our transmission signal is a linear frequency sweeping chirp from $20kHz$–$23kHz$ with a sampling rate of $48kHz$, which is inaudible yet compatible with commodity smartphones. Each of the four speakers transmits the signal to distribute it uniformly through the space. The transmitter synchronizes using BLE with the mobile device while it records the room response. The synchronization error is less than $1ms$ where 95% is within $\pm 200\mu s$, which results in a ranging error of $\pm 6.8cm$. We based our transmitter design and time synchronization on the platform the authors described in [29]. In our version of the design, the speakers all transmit simultaneously (instead of cycling). We discuss the impact of ranging accuracy on system performance in section 5.

3.1 Image Source Model

In Figure 3, we show a block diagram of our algorithm's main components. The system starts by periodically transmitting acoustic signals into a room with a loudspeaker, while the user captures echoes reflected back from the surfaces at multiple locations. By the nature of sound propagation, the relative position between the speaker, the sampling locations, and the surrounding surfaces are embedded in the arrival time of the echoes. To model the echo propagation in a room, we assume the room to be a K-faced convex polyhedron and we adopt the image source (IS) model as described in [1]. The main principle of the IS model is to replace a reflection path from a real source by a direct path from an image source. Assuming the location of the source is known and the echoes obey the law of reflection, the image sources are obtained by mirroring the real source to the surfaces. We refer to a received echo with n reflections as n^{th}-order echo and its corresponding image source an n^{th}-order image source. We show an example of a first and second-order image source in Figure 4. The IS model directly links the location of the image sources and the room geometry; knowing the location of an image source is equivalent to knowing the location of a surface. Using the IS model, we can convert the room geometry reconstruction problem into an image source localization problem where typical indoor localization techniques

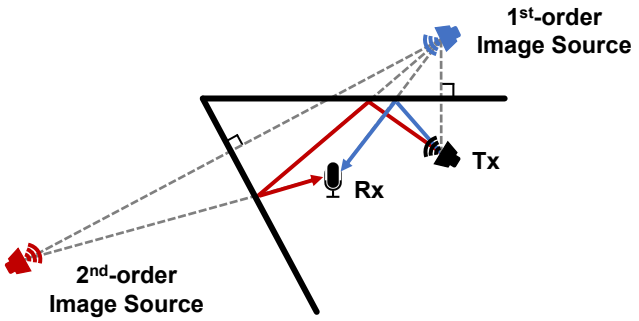


Figure 4: Illustration of the first-order and second-order image sources drawn with their reflection paths.

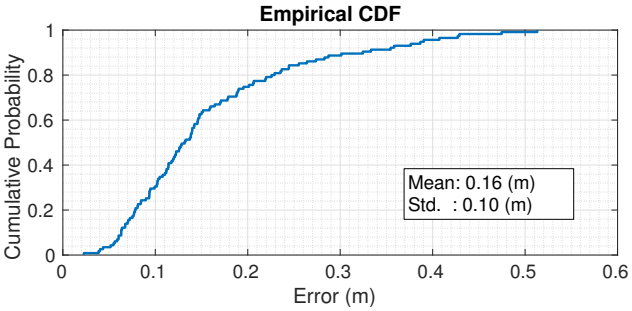


Figure 5: Cumulative distribution function of the localization error derived from ARKit traces.

can be applied. The main difference is that instead of localizing the real source using line-of-sight (LOS) signals, here we aim to localize multiple image sources simultaneously using the multipath reflections. To determine the location of an image source in 3D, we obtain ranging measurements to the image source from at least 4 different locations (more locations will improve performance). This is achieved by measuring the RIR from the received signal and convert the arrival time of the echoes into ranging estimates.

3.2 Visual Inertial Odometry for Localization

One of the critical enablers for being able to perform rich sonic sensing of environments is the ability to collect recordings at known locations rapidly. Recent advances in augmented reality (AR) [14, 23] have led to mobile phones that can precisely track their relative position over multiple meters using visual odometry (VO) fused with onboard inertial measurement (IMU) data. The so-called VIO systems track the motion of a field of feature points across image frames to accurately estimate the device's motion path. Apple and Google have released AR Kit and AR Core respectively that provide excellent VIO systems for mobile phones. Due to acoustic reciprocity, it is conceptually possible to swap microphone and speaker at any pairwise recording locations. Using a fixed speaker and any number of microphone sampling that can be localized moving through space, we can approximate arbitrarily dense sensing. In our prototype system, we use a single audio module to transmit the signal and record the data using the onboard microphone of a smartphone. The smartphone triggers the audio transmission over BLE and constantly streams back the recorded data to Matlab along with its location. We include a mount on the top of our speaker where a phone or tablet can be placed to maintain a constant origin coordinate. As we describe later in the evaluation section, this is useful for rendering objects like absorption heat-maps in their correct global coordinate frame for viewing with AR.

In our experiments, we used ARKit on an iPad pro running iOS 11 to evaluate the performance of currently available VIO systems on a mobile tablet. We collected ground-truth data at a set of 120 coordinates across our medium size room shown in Figure 13c by walking around the room while ARKit was streaming the tablet's coordinates at $30Hz$. Each time we reach a ground truth marker, we pressed a way-point button on the screen. Figure 5 shows a cumulative density function of the localization error after the phone had moved more than 100 meters over a 20 minute period. We see that the average error was $16cm$ with a worst-case error of only $50cm$. In section 5 we evaluate the impact of this performance on our ability to reconstruct room geometry and perform acoustic sensing. One can imagine that as AR systems evolve with the addition of depth cameras and higher resolution VIO, this performance will only continue to improve.

3.3 Acoustic Ranging

Pulse compression is a widely used technique in RADAR systems to improve the ranging resolution of signals. The transmitting signal, called a chirp, is a sinusoid where its frequency increases/decreases with time. When a chirp is cross-correlated with itself, the resulted signal behaves like a *sinc* signal around its peak value due to the frequency sweeping characteristic, and thus provides increased ranging resolution and signal-to-noise ratio (SNR). We selected a linear sweeping pattern due to its robustness against the Doppler shift. The ranging resolution of a linear frequency modulation chirp is inversely proportional to the sweeping bandwidth given by $\frac{c}{2\Delta f}$, where c is the speed of sound and Δf is the sweeping bandwidth. After pulse compression, the power of the received signal is increased by $T\Delta f$, where T is the pulse duration. In general, to achieve the best signal reception and resolution, one would select a chirp with long duration and large sweeping bandwidth. We chose a chirp length of $300ms$ based on the RT_{60} reverberation time of a typical size room [34]. Limiting the chirp length to the reverberation time helps to maximize the SNR of the received signal without spending excessive energy. Our chirp has a frequency sweeping range of $20kHz$–$23kHz$ such that it is inaudible to humans while capturing room geometry. We can later lower the chirp's frequency into the audible range for capturing sound absorption in the audible frequencies that are relevant to most acoustic engineers.

3.4 Preprocessing

In order to maximize the SNR of the received signal, we assume an additive white Gaussian noise (AWGN) model for the acoustic channel and apply a matched filter on the received signal. One side-effect of matched filtering a chirp signal is that it produces undesirable sidelobes around the main peaks, which makes peak detection difficult when multipath reflections are present. To reduce the effect of sidelobes, we apply an additional envelope detector on the matched signal. We search for the local maxima in the detected envelope, and then mapped them back to the nearest peaks in

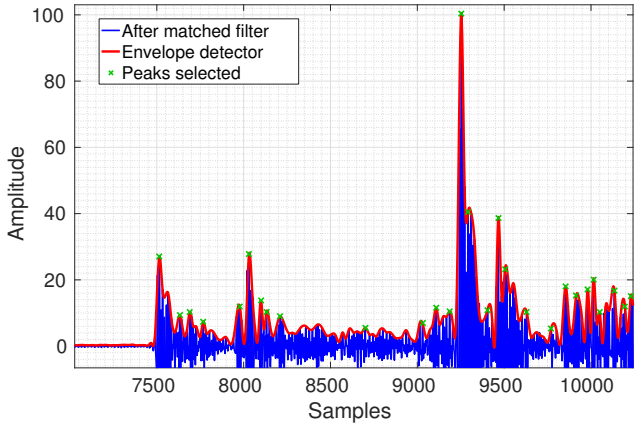


Figure 6: Example of the raw signal after matched filtering, the envelope detector and the selection of peaks.

the correlated signal. An example of the matched filtered signal, envelope detector, and the selection of peaks are shown in Figure 6.

4 GEOMETRY RECONSTRUCTION

In an ideal scenario, each microphone recording will capture echoes from all image sources, which means that a minimum of 4 measurements would be enough to localize all image sources simultaneously in 3D. However, this would require an ideal mapping between the echoes to the image sources that produce them, which is referred to as the echo labeling problem (subsection 4.2). The main challenge of echo labeling is that the arrival time of the echoes is location-dependent, and higher-order echoes from a surface may arrive earlier than the first-order echoes from another surface. One solution is to adopt the EDM formulation[11, 17] with a calibrated microphone array to filter out the incorrect echo combinations and reconstruct the room geometry. However, the EDM approach is not feasible when a self-tracking microphone is used to emulate a calibrated microphone array due to the different distances between microphone locations and inaccuracy in microphone tracking. In our algorithm, we use the same EDM formulation as a building block, but apply new optimization to address challenges in unconstrained microphone locations and missing/spurious echoes. We propose using an SDP method to improve the robustness against measurements uncertainty (subsection 4.3) and introduce combinatorial optimization to refine our solution (subsection 4.4).

Another challenge in unconstrained microphone locations is the exponential increase in computation complexity. To reduce the computation time, we propose a searching algorithm that utilizes the convexity of the EDM space to efficiently determine the candidate combinations (subsection 4.5). Once we have mapped the echoes to their image sources, we can localize them and reconstruct the room geometry. In reality, using the minimum number of measurements is often insufficient due to blind spots. Whether we can receive a reflection from a surface or not depends on both the measurement location and the geometry of reflective surfaces. In addition, each propagation path can be individually blocked or badly attenuated by clutter or people and may not be captured by the microphone. To deal with missing and spurious echoes, we divide all microphone locations into subsets and derive sub-optimal solutions accordingly. We later consolidate the sub-optimal solutions using a clustering algorithm with geometric properties to precisely reconstruct the room geometry (subsection 4.6).

4.1 Preliminaries

We introduce several notations to help formulating our approach and characterizing the Euclidean distance matrices (EDMs) using semi-definite matrices. We denote $\mathcal{S}^n$ as the space of $n \times n$ symmetric matrices and $\mathcal{E}^n$ as the space of $n \times n$ EDMs. A Euclidean distance matrix $D \in \mathcal{E}^n$ is defined by a set of n points $p_1, ..., p_n \in \mathbb{R}^r$ where

$$D_{ij} = \left\| p_i - p_j \right\|_2^2, \ \forall i, j = 1, ..., n \tag{1}$$

Let $\mathcal{S}_+^n$ denotes the cone of positive semi-definite matrices in $\mathcal{S}^n$ and we induce Löewner partial order $A \succeq B$ if $A - B \in \mathcal{S}_+^n$. We further denote the *hollow space* $\mathcal{S}_H^n := \{Y \in \mathcal{S}^n : \text{diag}(Y) = 0\}$ and the *centered space* $\mathcal{S}_C^n := \{Y \in \mathcal{S}^n : Ye = 0\}$, where $\text{diag}(\cdot)$ is the operator taking the diagonal elements of a matrix and $e \in \mathbb{R}^n$ is the vector of ones.

4.2 Echo Labeling and EDM

Correctly labeling the echoes to their reflected surfaces is the key to recovering the location of the image sources. At first, we assume an ideal scenario where measurements are precise and all reflections are received to show how to correctly label echoes. We later discuss how to relax these assumptions to deal with noisy measurement (subsection 4.3) and spurious/missing echoes (subsection 4.6).

While each microphone recording is composed of mixed echos from all image sources, the echo distances from every image source to the microphone locations are constrained by the relative positioning of the microphone locations. The EDM formulation therefore provides a naturally way to capture these constraints and helps to sort out the echo labeling problem. Assume we have K image sources in total and we collect echoes over N locations with known coordinates, we denote their corresponding Time-of-Flight (ToF) distances at each location as $d_n = [d_{n,1}, ..., d_{n,K}]$, $n = 1, ..., N$. By the definition of EDM, we can form a microphone EDM $D_{mic} \in \mathcal{E}^N$ using the pair-wise distances between the microphone locations. Similarly, if we introduce an additional image source to the set of microphone locations, we can form an augmented EDM based on D_{mic}, but with N unique echo distances added to the matrix. That is, for each image source k, there exists exactly one echo combination of squared distances, denote by $c_k = [c_{k,1}^2, ..., c_{k,N}^2]$ for $c_{k,n} \in d_n$, such that the augmented matrix $\bar{D}_k$ given by

$$\bar{D}_k = \begin{pmatrix} [D_{mic}] & \begin{bmatrix} c_k^T \end{bmatrix} \\ [c_k] & \mathbf{0} \end{pmatrix}$$

is also an EDM in $\mathcal{E}^{N+1}$. For the rest of the paper, such echo combination that corresponds to the same image source is referred to a *good* combination, or otherwise a *bad* combination. With this EDM formulation, we can solve the labeling problem by brute forcing all possible combinations and verifying whether the augmented matrix is an EDM or not. In practice, however, binary verification of EDM is uninformative because any augmented EDM will unlikely be a real EDM due to ranging error and numerical inaccuracy. Our goal is then redirected to finding the augmented matrices that are

the closest to real EDMs, which can be achieved by solving the Nearest EDM (NEDM) problem discussed in subsection 4.3.

4.3 Nearest EDM Problem

The delta between an augmented matrix and its NEDM often reflects the goodness of a combination, where a good combination typically has a small delta value. However, to ensure a small delta can only be derived from a good combination, precise estimates of microphone locations are required to recover from noisy measurements. In addition, since the total combinations grow exponentially with the number of microphone locations and the number of echo distances extracted per location, the microphone locations need to be close enough together (e.g. using a microphone array) to effectively reduce the number of feasible combinations and determine a unique solution [11]. In this paper, we relax these constraints and allow users to take measurements at arbitrary locations tracked by VIO. Our approach can effectively extend the distances between measurement locations, and therefore improve the accuracy of surface localization due to geometric dilution of precision (GDOP) [25, 35]. However, utilizing VIO also introduces a trade-off of additional localization errors and exponentially increased problem space. Therefore, a robust approach to solving NEDM problems is vital to the uniqueness and correctness of our solution.

Solving the NEDM problems with a low-rank constraint has been studied in several literature and known to be a non-convex optimization problem. Most current approaches rely on approximation methods. One popular approach to solve the NEDM problem is classical multidimensional scaling (cMDS), which has shown to be efficient in computation complexity. The cMDS starts by performing double centering on the augmented matrix and then projects the data into lower dimensions using the leading principal components. This method, however, inherits similar drawbacks of principal component analysis (PCA) in terms of being sensitive to outliers. In addition, cMDS has several features that are undesirable when dealing with noisy data [7]. Instead of directly projecting a target matrix onto $\mathcal{E}^n$ to find its closest approximation, cMDS projects it onto the cone of $\mathcal{S}_+^n$ and map it back to $\mathcal{E}^n$. This indirect mapping process makes the dissimilarities between the two matrices intractable and causes the result to be less robust. More generalized variations of MDS rely on distance scaling and direct approximation of target distances by minimizing stress based cost functions. However, these iterative approaches do not guarantee an optimal solution especially when input distances are noisy.

In this paper, we adopt the SDP approach as we find it to be more robust against noisy measurements in practice. EDM-based problems can be mathematically transformed into SDP formulations by leveraging the close relationship between EDMs and semi-definite matrices [9, 18–20, 28]. Given an EDM $D \in \mathcal{E}^n$, we can rewrite Equation 1 as

$$\begin{aligned} D_{ij} &= p_i^T p_i + p_j^T p_j - 2p_i^T p_j \\ &= Y_{ii} + Y_{jj} - 2Y_{ij} \end{aligned}$$

where $Y = p^T p$ is the Gram matrix of the point set that realizes the EDM. Since the Gram matrix is positive semi-definite, we observe the linear transformation $\mathcal{K}$ that maps $\mathcal{S}_+^n$ onto $\mathcal{E}^n$ ($\mathcal{K}(\mathcal{S}_+^n) = \mathcal{E}^n$) given by

$$\mathcal{K}(Y) := \text{diag}(Y)e^T + e\,\text{diag}(Y)^T - 2Y \tag{2}$$

And reversely we can derive the Moore-Penrose generalized inverse $\mathcal{K}^\dagger$ of $\mathcal{K}$ such that $\mathcal{K}\mathcal{K}^\dagger\mathcal{K} = \mathcal{K}$ given by

$$\mathcal{K}^\dagger(D) = -\frac{1}{2}V\text{offDiag}(D)V \tag{3}$$

where $V := I - ee^T/n$ is the geometric centering matrix and $\text{offDiag}(D) := D - \text{Diag}(\text{diag}(D))$ denotes the orthogonal projection onto the hollow matrices. This is a well-known result for the sufficiency of EDMs originally presented by Schoenberg [33]:

$$D \in \mathcal{E}^n \iff \begin{cases} -VDV \in \mathcal{S}_+^n \\ D \in \mathcal{S}_H^n \end{cases} \tag{4}$$

From Equation 2 we see an important property of the linear transformation $\mathcal{K}$ that it is *translational invariant*, meaning that the EDMs realized by a point set P and its infinite translational symmetries P' will be equivalent. To force the mapping $\mathcal{K}$ and $\mathcal{K}^\dagger$ to be bijective and prevent ambiguous solutions, we restrict the transformation to subspace $\mathcal{S}_C^n$ and $\mathcal{S}_H^n$ respectively, and have $\mathcal{K} : \mathcal{S}_C^n \bigcap \mathcal{S}_+^n \to \mathcal{E}^n$ a bijection and $\mathcal{K}^\dagger : \mathcal{E}^n \to \mathcal{S}_C^n \bigcap \mathcal{S}_+^n$ is its inverse. This result provides a key insight explaining the mapping between the convex cone of $\mathcal{E}^n$ and $\mathcal{S}_+^n$. However, even though the two convex cones can be related, a direct mapping between the two sets does not exist under the same dimensionality. In order to prevent unbounded optimal solutions [9], we define the transformation $\mathcal{K}_V : \mathcal{S}^{n-1} \to \mathcal{S}^n$ as

$$\mathcal{K}_V(X) := \mathcal{K}(V_n X V_n^T) \tag{5}$$

where $V_n \in \mathcal{R}^{n\times n-1}$ is the full rank skinny matrix such that $V_n^T e = 0$. We then have $\mathcal{K}_V(\mathcal{S}_+^{n-1}) = \mathcal{E}^n$ and $V_n X V_n^T$ is the Gram matrix of the point set. Similar to Equation 4, we derive another reformulation for the sufficiency of EDMs:

$$D \in \mathcal{E}^n \iff \begin{cases} -V_n^T D V_n \in \mathcal{S}_+^{n-1} \\ D \in \mathcal{S}_H^n \end{cases} \tag{6}$$

With Equation 6, we can rewrite the NEDM problem into a norm minimization problem that can be generally modeled as

$$\begin{aligned} \underset{X}{\text{argmin}} \quad & \left\| W \circ (\mathcal{K}_V(X) - \bar{D}) \right\|_F^2 \\ & (V_n X V_n^T)e = 0, \\ & X \succeq 0 \end{aligned} \tag{7}$$

where W is the weighted matrix that reflects the accuracy of the data, $\circ$ is the Hadamard product, and $\bar{D}$ is the target matrix we want to approximate. We drop the hard rank constraint as a relaxation to prevent non-convexity. Also note that X is solved in a lower dimension and we can recover the optimal distance matrix using $\mathcal{K}_V(X)$. We choose the Frobenius norm for our objective function since it naturally connects with the Euclidean distance space and it is strictly convex. We select the optimizer MOSEK [2] for solving the SDP problem and the combinatorial optimization problem next discussed in subsection 4.4.

4.4 Combinatorial Optimization

By solving the NEDM problem, we are able to score all echo combinations based on their augmented matrices' proximity to the closest EDM. We determine their scores to be inversely proportional to

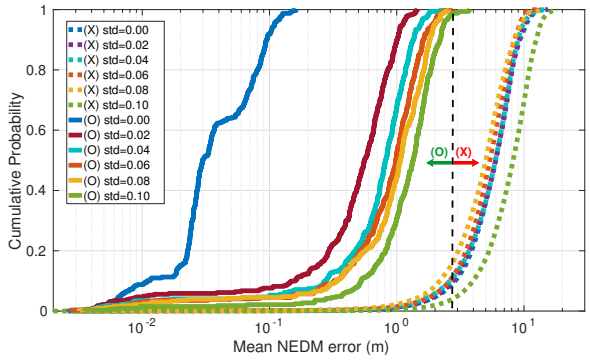


Figure 7: Mean NEDM error ($\mathcal{E}^5$) compared between the good (O) and the bad (X) combinations with ranging errors drawn from a Gaussian distribution.

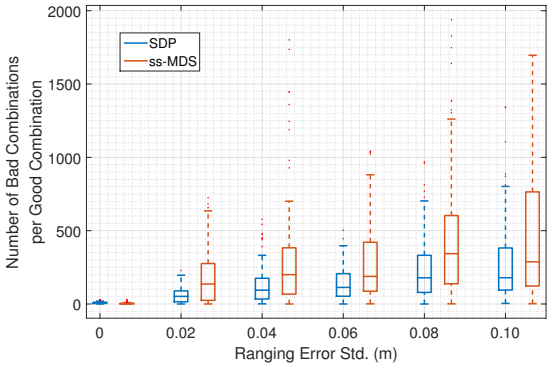


Figure 8: The average number of bad combinations with lower NEDM error increases with ranging error.

their NEDM error reported in Equation 7. In an ideal scenario, we could find every good combination by selecting the one with the highest score. In practice, however, a random bad combination could potentially produce a lower NEDM error due to noisy measurements and inaccuracy in the NEDM approximation, which eventually leads to erroneous image source locations.

To quantify the reliability of our NEDM approach in random room geometries with noisy measurements, we ran simulations to compare the distribution of the NEDM error between the good and the bad combinations. The results shown in Figure 7 indicates that good combinations typically have an error below a meter, while bad combinations yield much higher error over a wider distribution. Despite their distinct error distribution, bad combinations are orders of magnitude more than the good combinations that fall into the same error percentile. In Figure 8, we show mini-benchmarks comparing the performance between SDP and MDS on solving the NEDM problems with noise. While SDP achieves a lower number of bad combinations over a good combination, we observed hundreds of ambiguous solutions even with a small ranging error deviation of $4cm$. In order to refine our solution, we expand our objective function to minimize the total NEDM error among multiple combinations while limiting the occurrence of each echo distance in total. This, in turn, enforces constraints on our distance selection between combinations and greatly improves the chances of finding good combinations. The optimal selection is to find the set of combinations such that their combined score is the highest while satisfying

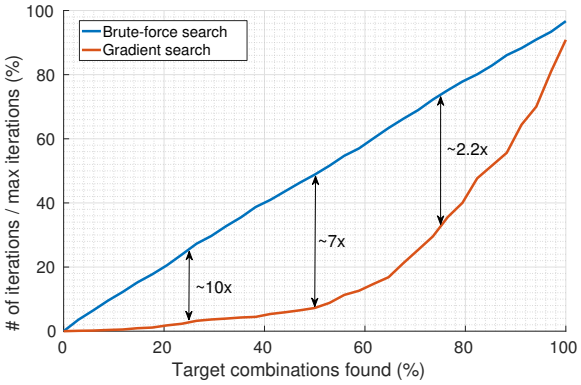


Figure 9: Improved performance with gradient-based local search compared to brute-force search.

all of the constraints, which can be formulated as a combinatorial optimization problem. Suppose we compute the score s_i for each echo combination c_i by solving $s_i = \text{NEDM}(D_{mic}, c_i)$, then we can solve the following combinatorial optimization problem using mixed integer programming in the form of

$$\begin{aligned} \max \quad & s^T x \\ \text{subject to} \quad & A^T x \le b \\ & x \in \{0,1\}^n \end{aligned}$$

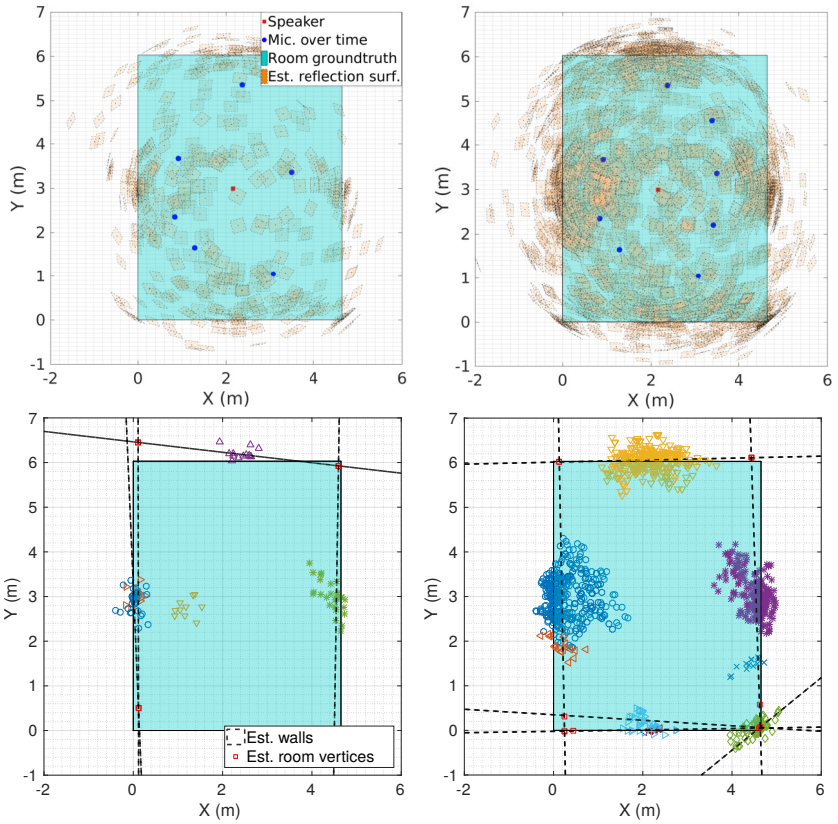
where s is the vector of scores derived from solving NEDM problem, x is the binary vector indicating whether a combination is selected, b is the vector for constraining the occurrence, and A is the constraint matrix that limits the selection between combinations given by

$$A_{ij} = \begin{cases} 1, & \text{if } d_j \in c_i \\ 0, & \text{otherwise} \end{cases} \quad \forall i = 1, ..., |c|,\ j = 1, ..., N$$

where $|c|$ is the total number of echo combinations. In our experiment, b is set as a vector of ones to achieve the best performance, but ideally, the constraint can be more relaxed when ranging resolution is low and peaks are inseparable due to close arrival times. Since the total number of surfaces is unknown, we encourage the algorithm to find as many surfaces as possible by pruning the combinations using an error threshold and solve the combinatorial optimization problem with an objective function that maximizes the total score. This error threshold helps to prevent overfitting and can be determined by simple heuristics or based on a prior estimation of the ranging error. While this greedy approach may allow some bad combinations to sneak through, it greatly improves the discovery of good combinations and benefits the following clustering algorithm (subsection 4.6) and overall performance. The objective function is biased toward combinations with low error due to the non-linearity of the inverse proportion operation when calculating the scores. The intuition is to increase the likelihood of selecting good combinations since the ratio of bad combinations over the good combinations decreases with NEDM error (Figure 7).

4.5 Gradient Search

Solving NEDM for each combination inevitably becomes a computational bottleneck due to the large combination space. In situations where the computation power is limited and/or the application is



(a) 6 microphone locations (b) 8 microphone locations

Figure 10: Top-down view of the clustering process in 3D. The detected surfaces increase exponentially with measurements, improving the clustering accuracy and overall reconstruction accuracy.

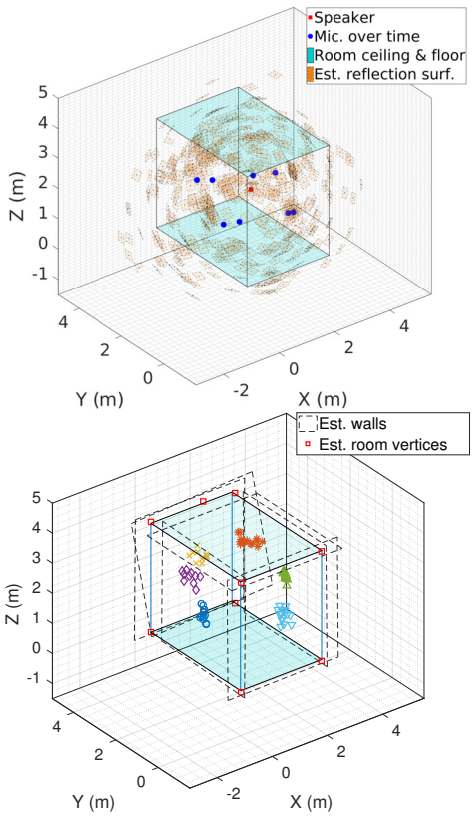


Figure 11: 3D view of the reconstruction rendered with the clustering result.

time sensitive, this impacts the reconstruction performance since we have to be more conservative about peak selections.

To reduce the computation time, we implemented an iterative gradient-based heuristic to directly search for combinations with low NEDM error. Since the problem is essentially a combinatorial search, our goal is to find the majority of the target combinations as a trade-off for computational efficiency. Similar to iterative local search (ILS) that is widely used for combinatorial problems, our algorithm dynamically refines its search direction by exploring neighborhood candidates of the best current solution. We exploit the convexity of EDM space and design our neighborhood function based on gradient descent. The search starts by randomly selecting a combination and solving the NEDM problem as described in Equation 7. At iteration t, we denote the selected combination as c_t and the resulting true EDM as $\mathcal{K}_V(X_t)$. To select the next combination c_{t+1}, we exploit $\mathcal{K}_V(X_t)$ and find the combination such that the gradient of the objective function with respect to the new augmented matrix $\bar{D}_{t+1}$ is closest to zero. Although the gradient does not guarantee an optimal searching direction, it can be derived in a computationally efficient manner and provides a reasonable approximation. Given our objective function f, the gradient is given as

$$\nabla_{\bar{D}} f = -2(W \circ W \circ (\mathcal{K}_V(X) - \bar{D}))$$

During the iterative search, we keep a history of the visited combinations and restart randomly when the gradient leads to a previously visited combination to improve the exploration of our search. Whenever the gradient reaches a local minimum and the NEDM error is below the given error threshold, we solve the combinatorial optimization previously mentioned in subsection 4.4 to dynamically trim the search space to increase diversity. These constraints are removed when there are no feasible combinations left and the search restarts with the history of the visited combinations. The algorithm ends when the certain percentage of total iterations is reached or a number of target combinations are found.

Since it is difficult to theoretically analyze the computation complexity for combinatorial search problems, we ran simulations to evaluate our heuristic. In Figure 9, we show a benchmark of our heuristic compared to brute force search in terms of the number of iterations spent. The proposed heuristic achieves 10 times faster at finding 25% of the target combinations and 7 times faster at the 50% mark. In general, the percentage of target combinations required for reconstruction could vary based on the ranging accuracy and the number of measurements. Experimentally, we find that 50% of the combinations are good enough to properly reconstruct the room geometry.

4.6 Reflective Surface Estimation

In order to localize an unknown number of reflective surfaces within a reasonable amount of computation time, we accept sub-optimal solutions. Missing/spurious echoes from clutter, such as furniture and occupants, may also introduce bad image sources. The final step of our algorithm is to eliminate these outliers and improve the localization of the good image sources. This is possible since that bad image sources would be scattered in space due to the randomness of the combinations, while the good ones would converge at their true locations due to matching echo distances from multiple locations. As shown in Figure 10, their corresponding surfaces, which are the bisectors of image sources and the speaker location, would also follow the same pattern.

To minimize the impact of clutter reflections, we recover the locations by selecting 4 random microphone locations at a time (minimum for 3D reconstruction) and iterate through different combinations of microphone locations. During this process we apply clustering on the combined results until the desired number of clusters are found. The algorithm therefore discovers larger reflectors (walls) first since they provide more consistent reflections and form clusters faster. To discover smaller reflectors, we can iteratively remove the processed clusters and continue the clustering process with wider radius. In general, the resolution of features the system can detect as surfaces is a function of the number of measurements and compute time. Capturing the first dozen major features is quite feasible, but the complexity increases quickly for higher resolution maps. In this paper, we chose a density-based clustering algorithm DBSCAN [21] due to its robustness to outliers and zero prior knowledge of the number of clusters or density distribution. One drawback of the DBSCAN algorithm is that the clustering results are sensitive to the minimum neighborhood points and neighbor distance. Through experiments we find the results to be the most stable when the neighborhood point is set to three times the reconstruction dimensionality, and the neighbor distance is determined based on our estimation of the ranging error. Once the clusters are found, each corresponding surface is determined as the plane passes through the geometrical center of the cluster with a normal vector pointing toward the speaker.

While clustering copes with the problems of missing echoes and having an unknown number of surfaces, the algorithm also captures the clusters from higher-order reflections. In order to bypass the process of identifying and eliminating higher-order image sources, we observe in Figure 12 that a virtual surface generated by a second-order echo will always cross their intersection. If the two reflected surfaces are adjacent to each other, the intersection will be an edge of the room; if they are not adjacent, the intersection will be distant from the room polyhedron. In fact, this result can be mathematically proven using geometry and holds for both 2D and 3D scenarios. Our algorithm leverages this geometry property and determines the room geometry as the smallest convex polyhedron within the virtual surfaces that bounds all of the microphone locations. As a result, the effects of second-order reflections and missing first-order reflections on the reconstructed room geometry is minimized in the presence of noise.

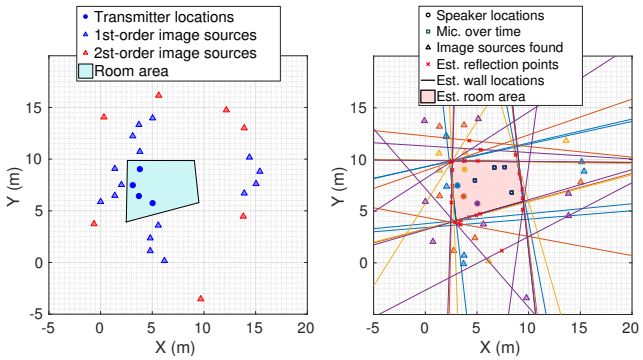


Figure 12: (Left) Room geometry with all feasible first and second-order image sources. (Right) Overlays of the reconstruction results where each is computed using data from a single speaker. The geometry is determined as the smallest convex polyhedron that bounds the microphone locations.

5 EVALUATION

In this section, we first evaluate the system's performance in a series of simulated environments. We then experimentally validate the results in a variety of rooms with real phone recordings.

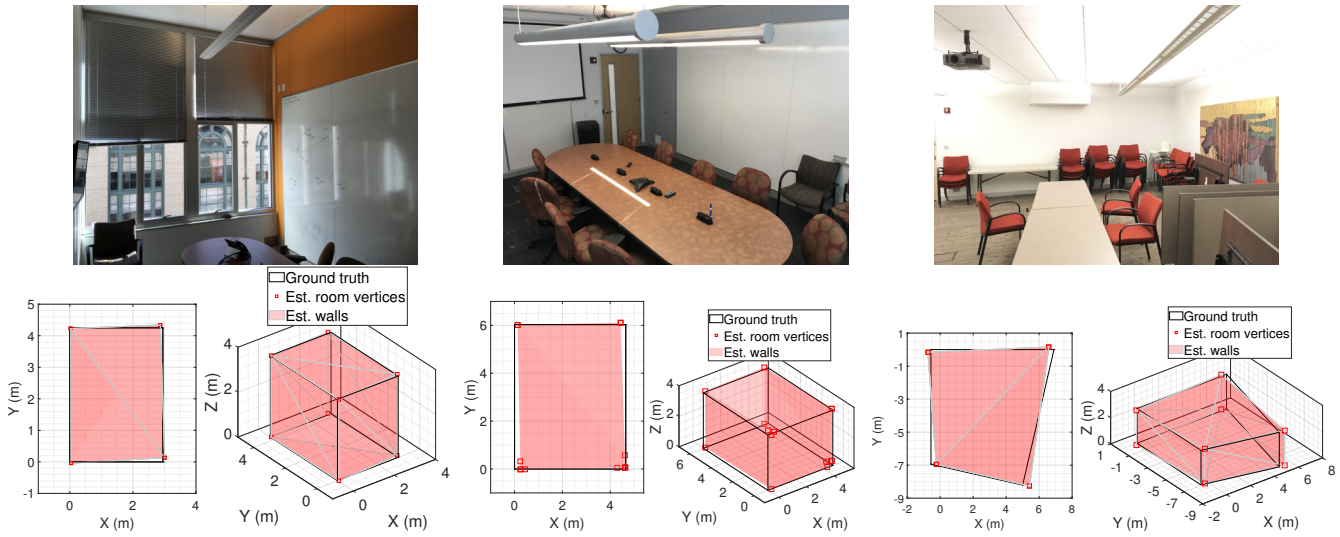
5.1 Simulation

To evaluate the impact of ranging error and the positioning between the room, speaker and microphone locations, we ran simulations with a set of random room geometries with 6–8 walls length from 5–10m with a minimum angle of 30°. Speaker and microphone locations are randomly selected with at least 50cm away from the walls. We used ray tracing to validate first and second-order image sources and to simulate path loss. Higher order reflections are dropped since they are rarely detected in reality due to attenuation. We added additional ranging bias and set the sound pressure level (SPL) to 65dB at 1 meter consistent with readily available commercial hardware. The wall absorption coefficient is set to 0.5 to simulate the absorption of common materials in the chirp's sweeping frequency [10]. For each parameter configuration, we ran at least 20 simulations. Since the estimated room geometry is translation and rotation invariant, we find the optimal rotation matrix that minimizes the root mean squared (RMS) error on microphone locations to align the result with the global coordinates system for visualization.

To measure the similarity between the ground truth and our estimation of the room, namely polyhedron A and B, we use the following criteria based on their overlapping volume and union volume, given by

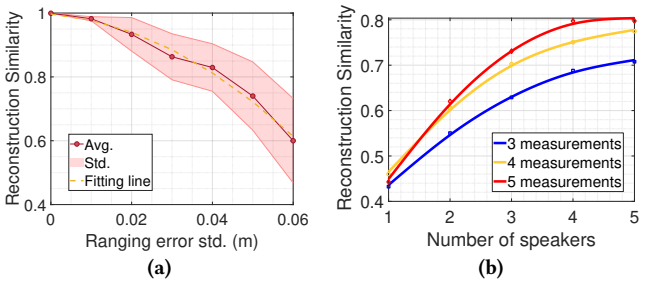
$$\text{Similarity} = \frac{A \cap B}{A \cup B} \tag{8}$$

The similarity metric is strict since it reflects not only the ranging error but also captures the translation, orientation and scaling for each surface. When computing the similarity in cases where walls are missing and the estimated polyhedron is not bounded, we artificially added the ground truth wall so the similarity can be determined, but penalize it by the percentage of added walls to the total number of walls. If the estimated polyhedron is bounded despite missing walls, the same rule is applied when it results in a better similarity to ensure a fair comparison. In Figure 12 we show



(a) Small conference room (b) Medium size conference room (c) Medium size breakout room

Figure 13: Experiment environments (Top) and their 3D reconstruction over ground truth (Bottom)



(a) (b)

Figure 14: Simulated reconstruction similarity with (a) varying ranging errors and (b) varying number of speakers and microphone measurements.

Room Size	Max. Similarity			Min. # of Locations		
	MDS [11]	MDS$^+$	SDP$^+$	≥70%	≥80%	≥90%
(a) 60.6 (m^3)	48.2%	63.5%	94.6%	6*	6*	7*
(b) 103.7 (m^3)	50.8%	68.4%	90.9%	6	8*	9*
(c) 132.2 (m^3)	46.0%	61.1%	90.5%	9	10*	11*

Table 1: System performance and minimum microphone locations to achieve certain similarity threshold with SDP$^+$ ($^+$ with proposed optimization, * all walls are discovered).

a simulated reconstruction of a room with a ranging error std of 3cm and achieve a similarity of 93%.

In Figure 14a, we show the overall reconstruction similarity with ranging error drawn from a normal Gaussian distribution with varying standard deviations. The similarity is found to be sensitive to the ranging accuracy and drops quickly once the std passes 4cm. Still, we are able to achieve 75% reconstruction accuracy on average with an std of 5cm. The relative positioning between the room, speaker and microphone also have an impact on the reconstruction accuracy. Most of the reconstruction error is caused by echoes arriving closely in time, where the algorithm fails to isolate the peaks or falsely selects the wrong echo combinations. To prevent simultaneous echo arrivals, the microphone locations should avoid lying on the symmetric axes of the room geometry. The algorithm can also be extended to accommodate multiple speakers by combining results during the clustering phase. In Figure 14b we showed the worst-case reconstruction similarity with varying number of speakers and microphone locations. With diverse speaker and microphone locations, we can effectively avoid scenarios where image sources cannot be localized due to geometry constraint. Diminishing return is observed when every wall is localizable and adding more measurements only improves slightly on estimation accuracy.

5.2 Real Environments

In Figure 13 we show photographs of three experimental environments. Two are small- and medium-size rooms with a shoe-box shape and the third is a larger room with an irregular polygon. We assume that blank walls are acoustically reflective and hence that room wall geometry is an accurate proxy for acoustic reflection. In our experiments, we placed the speaker close to the center of the room to provide better coverage and collected data at 12 random microphone locations in the presence of common obstacles (e.g. tables and chairs). Below each photograph, we show its reconstructed room geometry overlay on top of the ground truth. While we reconstructed the room geometry using location estimates provided by VIO, we also ground truth the locations to further evaluate the impact of localization error in post-processing.

An analytic comparison between the proposed method with related work is not trivial since most other approaches have vastly different assumptions and evaluation metrics. In Table 2 we summarize the assumptions and limitations of our proposed approach compared to related work. We believe the most relevant direct comparison can be made with [11] since it shares a similar problem formulation. We estimated room geometry using the same collected dataset and preprocessing tool but applied different optimization techniques accordingly. As shown in Table 1, the reconstruction with [11] performs poorly with an average around 50% similarity even when all 12 measurements are used. With the proposed optimization applied, we find the reconstruction using MDS-based

Summary	Proposed	Dokmanić et al.[11]	Jager et al.[17]	Moore et al.[24]	Zhou et al.[38]
# of speaker(s)	1	1	2	1	1
# of mic.(s)	1	5	5	1	2
Synchronized tx/rx	Yes	Yes	Yes	No	Yes
Known # of walls	No	Yes	Yes	Yes	No
Receive all 1^{st} order echoes	No	Yes	Yes	Yes	Yes
Method	EDM, SDP, combinatorial optimization, clustering, geometry properties, VIO	EDM, MDS	EDM, MDS, graph theory	Geometry properties	IMU, measurement gestures
Complexity	High	High	Medium	Medium	Low
Evaluation environment	Real-world	Real-world	Simulation	Simulation	Real-world
Cons	Require localization of the receiver.	Require careful calibration of the microphone array.	Require multiple speakers and microphones.	Assume 2D rectangular room shape.	Require localization of the receiver and additional user effort. Limited sensing range.

Table 2: System comparison with related work

approaches to be less ideal (65%) and the robustness in the NEDM approximation highly impact the clustering results and overall reconstruction accuracy. In comparison, our approach is able to achieve more than 90% reconstruction similarity using a maximum of 11 microphone locations. Increased number of microphone locations can effectively reduce the impact of noise and missing echoes which in turn improves reconstruction accuracy.

Across different rooms, the number of measurements required to achieve the same level of similarity slightly increases with the size of the room and the number of walls. The performance loss is mainly caused by signal attenuation and shadowing from obstacles. In our current implementation, the system can support a maximum sensing range of $10m$ when the speaker, reflection surface, and microphone are within LOS. The ranging accuracy and overall performance degrade when more obstacles are introduced to block the signal, but it can be compensated by increased output power and increased number of measurements from diverse locations. As the number of walls increases, more measurement locations are often required to prevent blind spots and disambiguate simultaneous echo arrivals from multiple walls. This phenomenon is remarkably noticeable when reconstructing a curved wall where its surface can be treated as numerous small facets. In most scenarios, a curved wall is reconstructed as its first-order or second-order polygonal approximation due to ranging inaccuracy and the clustering algorithm. However, with more sophisticated signal processing tools, it is possible to improve the ranging resolution and achieve finer approximation of the curved surfaces. The convex restriction on the room geometry is imposed simply to reduce the area of potential blind spots. In theory, our algorithm also works with concave room geometry as long as a minimum of 4 echoes can be received from each wall, or else only a partial geometry would be reconstructed.

Finally, while the reconstruction performance is evaluated using VIO readings, we also collected ground truth of the measurement locations as a way to evaluate the impact of localization error in post-processing. In Figure 15, we find the performance starts degrading sharply once the localization error exceeds $0.2m$, but appears robust to the typical levels of noise we see from ARKit traces ($0.16m$) and aligns with our empirical results.

5.3 AR Demonstration App

As a way to demonstrate the effectiveness of our acoustic sensing approach, we developed an AR phone application that can visualize absorption on surfaces in a room. We ran Synesthesia in a small room and collected data from 20 microphone locations with a series

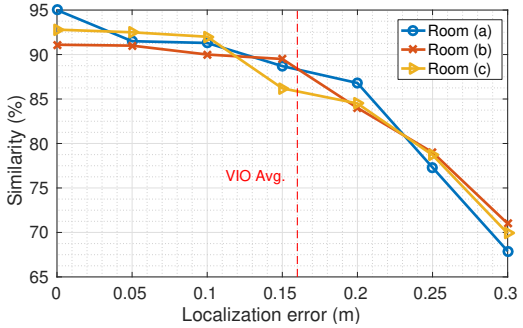


Figure 15: Impact of localization error to reconstruction similarity. VIO average highlighted.

of sound absorbing pads hanging on a wall with one removed to increase reflectivity. Once the model of the room was reconstructed, we simultaneously learned the locations of the echo reflections along with their propagation paths. We then computed the combined frequency response of the speaker and microphone using the intensity of the received LOS signal. Finally, we approximated each reflection surface's absorption coefficient based on the intensity of the reflected signal window and its estimated propagation path. To ensure a constant global coordinate system, we started the AR session mounting the phone on top of the speaker.

Figure 16 shows a photo of the AR app running where the colored circles are registered in 3D AR to visualize the sound absorption. Each color is mapped to the absorption coefficient across a particular frequency range. While each microphone measurement gives only one absorption estimate on each wall, the mobile platform allows the user to quickly capture multiple measurements and create a dense absorption map of the environment. Note that once we have obtained the room model, determining the locations of new surfaces and their absorption coefficients can be computed rather efficiently.

5.4 Limitation and Future Work

One of the main drawbacks of our current system is its long computation time. To achieve the best reconstruction similarity, it takes on average 1 hour in Matlab with dual Intel Core i7 CPUs. However, we believe with an optimized implementation, parallelization, and GPU acceleration, the computation time could be reduced to minutes, and further reduced with partial prior knowledge on the reflector or microphone locations. Additional speedup would likely involve improvements in peak selection and the searching algorithm. In most applications, we envision users capturing an image, pushing the data to the cloud and retrieving it later for viewing.

As future work, we intend to evaluate our system in more challenging environments and further study the impact of occupants and non-convex room geometries. When operating in large spaces, the system would also require a larger amplifier and transducer to support a proportionally increased output power. Alternatively, images from multiple smaller spaces could be stitched together.

6 CONCLUSION

We showed the feasibility of estimating the locations of reflective surfaces and forming a high-resolution image of the space, given a single acoustic source and multiple noisy microphone measurements from a mobile device. Our proposed algorithm utilizes a pipeline of optimization techniques to eliminate conventional assumptions on room geometry and detection of echoes. We showed through both simulation and experimentation that we are able to reconstruct the room geometry with more than 90% accuracy.

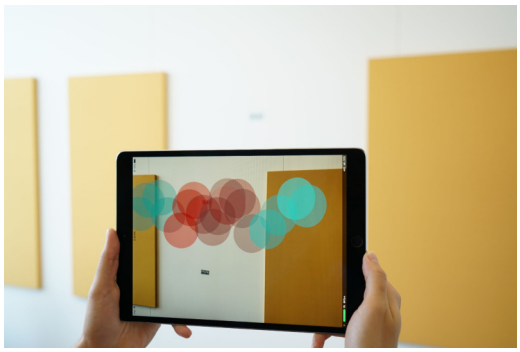


Figure 16: Visualization of sound absorption in AR. Red in color indicates less absorption while blue denotes more.

REFERENCES

[1] Jont B. Allen and David A. Berkley. 1979. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America* 65 (1979).

[2] Erling D. Andersen and Knud D. Andersen. 2000. *The Mosek Interior Point Optimizer for Linear Programming: An Implementation of the Homogeneous Algorithm.* Springer US, Boston, MA, 197–232.

[3] Fabio Antonacci, Jason Filos, Mark R. P. Thomas, Emanuël A. P. Habets, Augusto Sarti, Patrick A. Naylor, and Stefano Tubaro. 2012. Inference of Room Geometry From Acoustic Impulse Responses. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 10 (July 2012), 2683 – 2695.

[4] Google ARCore. 2019. https://developers.google.com/ar/. [Online; accessed 16-February-2019].

[5] Apple ARKit. 2019. https://developer.apple.com/arkit/. [Online; accessed 16-February-2019].

[6] A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher. 2014. Structured Sparsity Models for Reverberant Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, 3 (March 2014), 620–633.

[7] Lawrence Cayton and Sanjoy Dasgupta. 2006. Robust Euclidean Embedding. In *Proceedings of the 23rd International Conference on Machine Learning (ICML '06).* ACM, New York, NY, USA, 169–176.

[8] Marco Crocco, Andrea Trucco, Vittorio Murino, and Alessio Del Bue. 2014. Towards fully uncalibrated room reconstruction with sound. In *Signal Processing Conference (EUSIPCO), Proceedings of the 22nd European.*

[9] Jon Dattorro. 2005. Convex Optimization & Euclidean Distance Geometry.

[10] Andrzej Dobrucki, Bronislaw Zoltogorski, Piotr Pruchnicki, and Romuald Bolejko. 2010. Sound-Absorbing and Insulating Enclosures for Ultrasonic Range. *Archives of Acoustics* 35, 2 (May 2010), 157 – 164.

[11] Ivan Dokmanić, Reza Parhizkara, Andreas Walthera, Yue M. Lub, and Martin Vetterlia. 2013. Acoustic echoes reveal room shape. *Proceeding of the National Academy of Science of the United States of America* 110, 30 (July 2013).

[12] Amazon Echo. 2019. https://www.amazon.com/echo. [Online; accessed 16-February-2019].

[13] Yi Fang, Haihong Feng, and Youyuan Chen. 2018. A robust interaural time differences estimation and dereverberation algorithm based on the coherence function. *Applied Acoustics* 129 (2018), 126 – 134. http://www.sciencedirect.com/science/article/pii/S0003682X17302852

[14] Google Glass. 2018. https://www.x.company/glass/. [Online; accessed 13-March-2018].

[15] Google Home. 2019. https://store.google.com/product/google_home. [Online; accessed 16-February-2019].

[16] Apple Homepod. 2019. https://www.apple.com/homepod/. [Online; accessed 16-February-2019].

[17] Ingmar Jager, Richard Heusdens, and Nikolay D. Gaubitch. 2016. Room geometry estimation from acoustic echoes using graph-based echo labeling. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on.*

[18] Charles R. Johnson. 1995. Connections between the real positive semidefinite and distance matrix completion problems. *Linear Algebra Appl.* 223 - 224 (July 1995), 375 – 391.

[19] Nathan Krislock and Henry Wolkowicz. 2012. *Euclidean Distance Matrices and Applications.* Springer US, Boston, MA, 879–914.

[20] Monique Laurent. 1998. A connection between positive semidefinite and euclidean distance matrix completion problems. *Linear Algebra and its Application* 273, 1 - 3 (April 1998), 9 – 22.

[21] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *International Conference on Knowledge Discovery and Data Mining.*

[22] H. Malik. 2013. Acoustic Environment Identification and Its Applications to Audio Forensics. *IEEE Transactions on Information Forensics and Security* 8, 11 (Nov 2013), 1827–1837.

[23] Microsoft Hololens. 2018. https://www.microsoft.com/en-us/hololens/. [Online; accessed 13-March-2018].

[24] Alastair H. Moore, Mike Brookes, and Patrick A. Naylor. 2013. Room geometry estimation from a single channel acoustic impulse response. In *Signal Processing Conference (EUSIPCO), Proceedings of the 21st European.*

[25] Neal Patwari, Joshua N Ash, Spyros Kyperountas, Alfred O Hero III, Randolph L Moses, and Neiyer S Correal. 2005. Locating the nodes: cooperative localization in wireless sensor networks. *Signal Processing Magazine, IEEE* 22, 4 (2005), 54–69.

[26] F. Peng, T. Wang, and B. Chen. 2015. Room shape reconstruction with a single mobile acoustic sensor. In *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP).* 1116–1120.

[27] Marc Pollefeys and David Nister. 2008. Direct computation of sound and microphone locations from time-difference-of-arrival data. In *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on.*

[28] Hou-Duo Qi and Xiaoming Yuan. 2014. Computing the nearest Euclidean distance matrix with low embedding dimensions. *Mathematical Programming* 147, 1 (2014), 351–389.

[29] Niranjini Rajagopal, Patrick Lazik, Nuno Pereira, Sindhura Chayapathy, Bruno Sinopoli, and Anthony Rowe. 2018. Enhancing Indoor Smartphone Location Acquisition Using Floor Plans. In *Proceedings of the 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN '18).* IEEE Press, Piscataway, NJ, USA, 278–289. https://doi.org/10.1109/IPSN.2018.00056

[30] Tilak Rajapaksha, Xiaojun Qiu, Eva Cheng, and Ian Burnett. 2016. Geometrical room geometry estimation from room impulse responses. In *In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on.*

[31] L. Remaggi, P. J. B. Jackson, W. Wang, and J. A. Chambers. 2015. A 3D model for room boundary estimation. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 514–518.

[32] Wallace Clement Sabine. 1923. Collected Papers on Acoustics. *Harvard University Press* (1923).

[33] I. J. Schoenberg. 1935. Remarks to Maurice Frechet's Article "Sur La Definition Axiomatique D'Une Classe D'Espace Distances Vectoriellement Applicable Sur L'Espace De Hilbert. *Annals of Mathematics* 36, 3 (1935), 724–732.

[34] Oliver Shih and Anthony Rowe. 2015. Occupancy Estimation Using Ultrasonic Chirps. In *Proceedings of the ACM/IEEE Sixth International Conference on Cyber-Physical Systems (ICCPS '15).* ACM, New York, NY, USA, 149–158.

[35] Maurizio Spirito et al. 2001. On the accuracy of cellular mobile station location estimation. *Vehicular Technology, IEEE Transactions on* 50, 3 (2001), 674–685.

[36] Sakari Tervo and Timo Tossavainen. 2012. 3D room geometry estimation from measured impulse responses. In *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on.*

[37] Takuya Yoshioka, Armin Sehr, Marc Delcroix, Keisuke Kinoshita, Roland Maas, Tomohiro Nakatani, and Walter Kellermann. 2012. Making Machines Understand Us in Reverberant Rooms: Robustness Against Reverberation for Automatic Speech Recognition. *IEEE Signal Processing Magazine* 29 (2012), 114–126.

[38] Bing Zhou, Mohammed Elbadry, Ruipeng Gao, and Fan Ye. 2017. BatMapper: Acoustic Sensing Based Indoor Floor Plan Construction Using Smartphones. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '17).* ACM, New York, NY, USA, 42–55. https://doi.org/10.1145/3081333.3081363

[39] D. N. Zotkin, R. Duraiswami, and L. S. Davis. 2004. Rendering localized spatial audio in a virtual auditory space. *IEEE Transactions on Multimedia* 6, 4 (Aug 2004), 553–564.