Voice Localization Using Nearby Wall Reflections

Sheng Shen
University of Illinois at
Urbana-Champaign
sshen19@illinois.edu

Daguan Chen
University of Illinois at
Urbana-Champaign
daguanc2@illinois.edu

Yu-Lin Wei
University of Illinois at
Urbana-Champaign
yulinlw2@illinois.edu

Zhijian Yang
University of Illinois at
Urbana-Champaign
zhijian7@illinois.edu

Romit Roy Choudhury
University of Illinois at
Urbana-Champaign
croy@illinois.edu

ABSTRACT

Voice assistants such as Amazon Echo (Alexa) and Google Home use microphone arrays to estimate the angle of arrival (AoA) of the human voice. This paper focuses on adding user localization as a new capability to voice assistants. For any voice command, we desire Alexa to be able to localize the user inside the home. The core challenge is two-fold: (1) accurately estimating the AoAs of multipath echoes without the knowledge of the source signal, and (2) tracing back these AoAs to reverse triangulate the user's location.

We develop *VoLoc*, a system that proposes an *iterative align-and-cancel* algorithm for improved multipath AoA estimation, followed by an *error-minimization* technique to estimate the geometry of a nearby wall reflection. The AoAs and geometric parameters of the nearby wall are then fused to reveal the user's location. Under modest assumptions, we report localization accuracy of 0.44 m across different rooms, clutter, and user/microphone locations. *VoLoc* runs in near real-time but needs to hear around 15 voice commands before becoming operational.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing**; • **Hardware** → *Signal processing systems*; • **Information systems** → *Location based services.*

KEYWORDS

Amazon Alexa, smart home, voice assistant, source localization, microphone array, acoustic reverberation, angle-of-arrival, edge computing

ACM Reference Format:
Sheng Shen, Daguan Chen, Yu-Lin Wei, Zhijian Yang, and Romit Roy Choudhury. 2020. Voice Localization Using Nearby Wall Reflections. In *The 26th Annual International Conference on Mobile Computing and Networking (MobiCom '20), September 21–25, 2020, London, United Kingdom.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3372224.3380884

1 INTRODUCTION

Voice assistants such as Amazon Echo and Google Home continue to gain popularity with new "skills" being continuously added to them [6, 9, 22, 38, 60, 69]. A skill coming to Alexa is the ability to infer emotion and age group from the user's voice commands [6, 9, 38]. More of such skills are expected to roll out, aimed at improving the contextual background of the human's voice command. For instance, knowing a user's age may help in retrieving information from the web and personalizing human-machine conversations.

Towards enriching multiple dimensions of context-awareness, companies like Amazon, Google, and Samsung are also pursuing the problem of user localization [8, 12, 35, 43]. Location adds valuable context to the user's commands, allowing Alexa to resolve ambiguities. For instance: (1) Knowing the user's location could help determining which light the user is referring to, when she says "turn on the light". Naming and remembering every light (or fans, thermostats, TVs, and other IoT devices) is known to become a memory overload for the users [2, 5]. Similarly, Alexa could help energy saving in smart buildings if it understands where the user is when she says "increase the temperature". (2) More broadly, location could aid speech recognition by narrowing down the set of possible commands [4, 46, 62]. If Alexa localizes Jane to the laundry machine, then a poorly decoded command like "add *urgent* to groceries" could be correctly resolved to "*detergent*". In fact, Google is working on "generating kitchen-specific speech recognition models", when its voice assistant detects "utterances made in or near kitchens" from the user [31, 75]. (3) Lastly, localizing sounds other than voice – say footsteps or running water – could further enrich context-awareness [7]. Alexa could perhaps remind an independently living grandmother to take her medicine when she is walking by the medicine cabinet, or nudge a child when he runs out of the washroom without turning off the faucet.

These and other uses of location will emerge over time, and the corresponding privacy implications will also need attention. In this paper, however, we focus on exploring the technical viability of the problem. To this end, let us begin by intuitively understanding the general problem space, followed by the underlying challenges and opportunities.

The central question in voice-source localization is that an unknown source signal must be localized from a single (and small) microphone array. Relaxing either one of these two requirements brings up rich bodies of past work [24, 28, 37, 63, 66, 78, 79]. For instance, a known source signal (such as a training sequence or an

MobiCom '20, September 21–25, 2020, London, United Kingdom

ACM ISBN 978-1-4503-7085-1/20/09…$15.00
https://doi.org/10.1145/3372224.3380884

impulse sound) can be localized through channel estimation and fingerprinting [29, 59, 66, 79], while scattered microphone arrays permit triangulation [24, 25, 37, 63]. However, *VoLoc*'s aim to localize *arbitrary sound signals with a single device* essentially inherits the worst of both worlds.

In surveying the space of solutions, we observe the following: (1) Signal strength based approaches that estimate some form of distance are fragile due to indoor multipath. Amplitude variations across microphones are also small due to the small size of the microphone array. (2) Machine learning approaches to jointly infer the in-room reflections and per-user voice models seem extremely difficult, even if possible. Moreover, such training would impose a prohibitive burden on the users, making it unusable. (3) Perhaps a more viable idea is to leverage the rich body of work in *angle of arrival* (AoA). Briefly, AoA is the angular direction from which a signal arrives at a receiver. Voice assistants today already estimate the direct path's AoA and beamform towards the user [1, 3, 23, 36]. So one possibility is to detect additional AoAs for the multipath echoes and trace back the AoA directions to their point of intersection (via reverse triangulation).

While the idea of tracing back indoor multipath echoes (such as from walls and ceilings) for reverse triangulation is certainly not new [16], unfortunately, extracting the AoAs for individual echoes, especially indoors, is difficult even in today's state-of-the-art algorithms [41, 68]. Even the direct path AoA is often erroneous/biased in today's systems, and small AoA offsets magnify localization error. Finally, tracing back the AoAs requires the knowledge of the reflectors in the room, a somewhat impractical proposition. This is why existing work that leverage multipath reverse triangulation have assumed empty rooms, known sounds, and even near-field effects [16, 28, 59].

While the problem is non-trivial, application-specific opportunities exist:

- Perhaps not all AoAs are necessary; even two AoAs may suffice for reverse triangulation, so long as these AoAs are estimated with high accuracy. Of course, the reflector for the second AoA is still necessary.
- To connect to power outlets, Alexa is typically near a wall. If the AoA from the wall can be reliably discriminated from other echoes, and the wall's distance and orientation estimated from voice signals, then reverse triangulation may be feasible.
- Finally, the user's height can serve as an invariant, constraining the 3D location search space.

All in all, these opportunities may give us adequate ammunition to approach the problem. Thus, the core algorithmic questions boil down to *accurate AoA detection* and *joint wall geometry estimation.* These two modules form the technical crux of *VoLoc* – we discuss our core intuitions next.

■ **Accurate AoAs:** Accurate AoA estimation is difficult in multipath settings because each AoA needs to be extracted from a mixture of AoAs, caused by echoes. Existing algorithms try to align (beamform) towards different directions to find the energy maxima, but do not perform well because all the echoes are strongly correlated (elaborated in Section 2). We aim to break away from this approach, and our central idea is rooted in leveraging (1) slow velocity, and (2) *pauses* (or short silences) in acoustic signals. A voice command, for example, is preceded by silence. The ends of these silences are unique opportunities to observe the cascade of arriving signals, starting with the clean direct path first, followed by the first echo, second echo, and so on. This means that the direct path signal is essentially clean for a short time window, presenting an opportunity to accurately derive its AoA. Since the first echo is a delayed version of the direct path, this echo can be modeled and *cancelled* with appropriate alignment. This process can continue iteratively, and in principle, all AoAs and delays can be jointly extracted.

In practice, hardware noise becomes the limiting factor, hence cancellation errors accrue over time. Thus, *VoLoc* extracts accurate AoAs and delays for only the initial echoes and utilizes them for source localization.

■ **Wall Geometry Estimation:** Inferring source location from AoA requires geometric knowledge of signal reflectors. To cope with this requirement, existing work have assumed empty rooms with no furniture, and used non-acoustic sensors (such as cameras or depth sensors) to scan the walls and ceilings of the room [16]. Our opportunity arises from the fact that the wall near Alexa serves as a stable echo, i.e., it is always present. If the wall's distance and orientation can be estimated with respect to Alexa, then the echo's AoA and delay become a function of the user location. This also helps in discriminating the wall echo from other echoes, say from objects on the table around Alexa. The algorithmic challenge lies in estimating the wall's ⟨distance, orientation⟩ tuple from the same voice signals.

We address this problem by gathering signals from recent voice commands and asking the following question: *At what distance d and orientation θ must a reflector be, such that its echo arrives early and is frequently present in voice command signals?* We formulate this as an optimization problem with the error function modeled in terms of $\langle d, \theta \rangle$. This error is summed across multiple recent voice commands, and the minimum error yields the $\langle d, \theta \rangle$ estimates. We over-determine the system by fusing AoA, $\langle d, \theta \rangle$, and user height h, and converge to the user's indoor 2D location.

We implement *VoLoc* on an off-the-shelf hardware platform composed of a 6-microphone array, positioned in a circular shape like Amazon Echo (Figure 1). This was necessary to gain access to raw acoustic signals (commercially available Echo or Google platforms do not export the raw data). Our microphone array forwards the signal to a Raspberry Pi, which performs basic signal processing and outputs the data into a flash card, transmitted to our laptop over a WiFi direct interface. Experiment results span across AoA and location estimations in various environments, including student apartments, house kitchen, conference rooms, etc.

Our results reveal median localization accuracy of 0.44 m across a wide range of environments, including objects scattered around the microphone. In achieving this accuracy, the detected AoAs consistently outperform GCC-PHAT and MUSIC algorithms. *VoLoc* also estimates wall geometry (distance and orientation) with average accuracies of 1.2 cm and 1.4°, respectively. The results are robust across rooms, users, and microphone positions.

■ **Current Limitations:** We believe that blind voice-source localization remains a challenging problem in practice, and *VoLoc* addresses it under four geometric assumptions: (1) The user's height

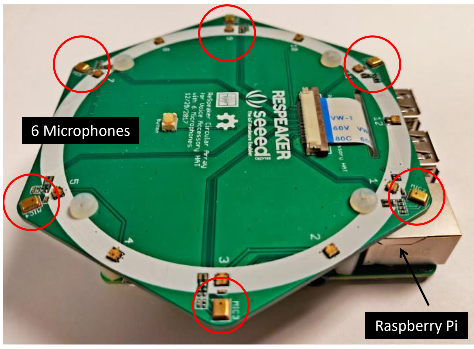


Figure 1: Seeed Studio off-the-shelf 6-microphone array, sitting on top of a Raspberry Pi.

is known. (2) The line-of-sight (LoS) path exists, meaning that obstructions between the user and the voice assistant do not completely block the signal. (3) The stable reflector is not too far away (so that its reflection is among the first few echoes). (4) The user is within $4-5$ m from the device (or else, slight AoA errors translate to large triangulation error). While future work would need to relax these assumptions, we believe the core AoA algorithm and the wall-geometry estimation in this paper offer an important step forward. To this end, our contributions may be summarized as:

- A novel *iterative align-and-cancel algorithm* that jointly extracts initial AoAs and delays from sound pauses. The technique is generalizable to other applications.
- An *error minimization formulation* that jointly estimates the geometry of a nearby reflector using only the recent voice signals.
- A *computationally efficient fusion* of AoA, wall-reflection, and height to infer indoor 2D human locations.

In the following, we expand on each of these contributions, starting with background on AoA.

2 BACKGROUND AND FORMULATION

This section presents relevant background for this paper, centered around array processing, angle of arrival (AoA), and triangulation. The background will lead into the technical problems and assumptions in *VoLoc*.

2.1 Array Processing and AoA

Figure 2 shows a simple 3-element linear microphone array with d distance separation. Assuming no multipath, the source signal $s(t)$ will arrive at each microphone as $x_1(t)$, $x_2(t)$ and $x_3(t)$, after traveling a distance of D_1, D_2 and D_3, respectively. Usually $\{D1, D2, D3\} \gg d$, hence these sound waves arrive almost in parallel (known as the far field scenario). From geometry, if the signal's incoming angle is θ, then the signal wave needs to travel an extra distance of $\Delta d = d\cos(\theta)$ to arrive at microphone M_2 compared to M_1, and an extra $2\Delta d$ at M_3 compared to M_1.

When the additional travel distance is converted to phase, the phase difference between $x_2(t)$ and $x_1(t)$ is $\Delta\phi = 2\pi d\cos(\theta)/\lambda$, and between $x_3(t)$ and $x_1(t)$ is $2\Delta\phi$. On the other hand, the amplitudes of $x_1(t)$, $x_2(t)$ and $x_3(t)$ will be almost the same, due to very minute differences among D_1, D_2 and D_3.[1] Thus, in general, the received signal vector can be represented as:

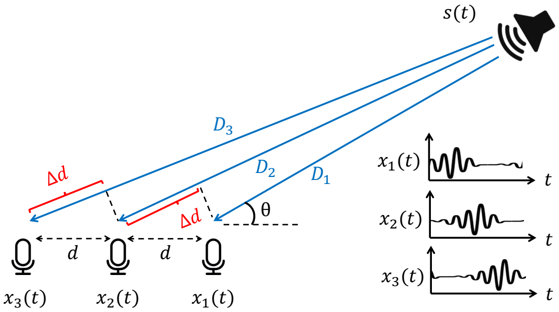


Figure 2: A simple 3-element microphone array.

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ x_1 e^{j\Delta\phi} \\ \vdots \\ x_1 e^{j(n-1)\Delta\phi} \end{bmatrix} = \begin{bmatrix} e^{j0} \\ e^{j\Delta\phi} \\ \vdots \\ e^{j(n-1)\Delta\phi} \end{bmatrix} x_1$$

■ **AoA Estimation without Multipath:** In reality, we do not know the signal's incoming angle θ, hence we perform AoA estimation. One solution is to consider every possible θ, compute the corresponding $\Delta\phi$, apply the appropriate negative phase shifts to each microphone, and add them up to see the total signal energy. The correct angle θ should present a maximum energy because the signals will be perfectly aligned, while others would be relatively weak. This AoA technique essentially has the effect of steering the array towards different directions of arrival, computing an AoA energy spectrum, and searching for the maximum peak. For a single sound source under no multipath, this reports the correct AoA direction.

■ **Impact of Multipath Echoes:** Now consider the source signal $s(t)$ reflecting on different surfaces and arriving with different delays from different directions. Each arriving direction is from a specific value of θ_i, translating to a corresponding phase difference $\Delta\phi_i$. Thus the received signal at each microphone (with respect to microphone M_1) is a sum of the same source signal, delayed by different phases. With k echoes, we can represent the received signal as:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} e^{j0} & e^{j0} & & e^{j0} \\ e^{j\Delta\phi_1} & e^{j\Delta\phi_2} & \dots & e^{j\Delta\phi_k} \\ \vdots & \vdots & & \vdots \\ e^{j(n-1)\Delta\phi_1} & e^{j(n-1)\Delta\phi_2} & \dots & e^{j(n-1)\Delta\phi_k} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_k \end{bmatrix}$$

■ **Estimating AoA under Multipath:** The earlier AoA technique (of searching and aligning across all possible θ_i) is no longer accurate since phase compensation for an incorrect θ_i may also exhibit strong energy in the AoA spectrum (due to many strongly correlated paths). Said differently, searching on θ_i is fundamentally a cross-correlation technique that degrades with lower SNR. Since

[1] Sound amplitude attenuates with $1/r$ where r is traveled distance. For two paths of r and $r+\Delta d$, the relative amplitude difference is $[1/r - 1/(r+\Delta d)]/(1/r) \approx \Delta d/r$. If Δd=2 cm and r=2 m, there would be a 1% amplitude difference.

any path's SNR reduces with increasing echoes, AoA estimation is unreliable.

While many AoA-variants have been proposed [18, 20, 27, 39, 41, 61, 64, 67, 74, 76, 77], most still rely on cross-correlation. The most popular is perhaps GCC-PHAT [18, 20, 32, 39, 76] which compensates for the amplitude variations across different frequencies by whitening the signal. The improvement is distinct but does not solve the root problem of inaccurate alignment. Subspace based algorithms (like MUSIC, ESPRIT, and their variants [61, 64, 67, 72, 74]) are also used, but they rely on the assumption that signal paths are uncorrelated or can be fully decorrelated. Multipath echoes exhibit strong correlation, leaving AoA estimation a still difficult problem.

2.2 Reverse Triangulation

Even if AoAs are estimated correctly, localization would require knowledge of reflectors in the environment to reverse triangulate (Figure 3). While some past work has scanned the environment with depth cameras to create 3D room models [16], this approach is largely impractical for real-world users.

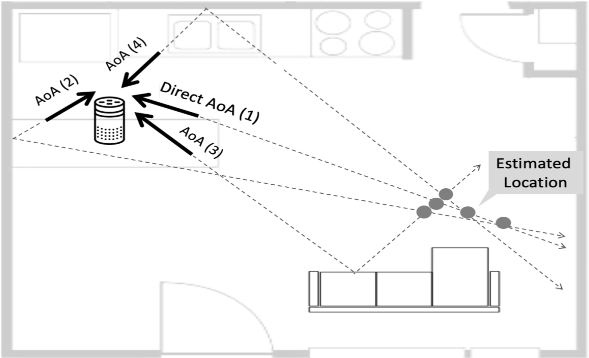


Figure 3: Reverse triangulation requires location of all reflector surfaces, making it impractical.

In principle, however, not all echoes are necessary for tracing back to the source location. The direct path's AoA and one other AoA would be adequate: say, AoA(1) and AoA(2) in Figure 3. Of course, the location and orientation of AoA(2)'s reflector still needs to be inferred. The authors of [28, 29] have attempted a related problem. They attempt to infer the shape of an empty room; however, they use precisely designed wideband signals, scattered microphone arrays, and essentially solve compute-intensive inverse problems [28, 29]. *VoLoc* takes on the simpler task of estimating one reflector position, but under the more challenging constraint of unknown voice signals and near real-time deadlines (a few seconds[2]).

2.3 Problem Statement and Assumptions

With this background, the problem in this paper can be stated as follows. Using a 6-microphone array, without any knowledge of the source signal, and under the following assumptions:

- Alexa located near a wall to connect to a power outlet
- User's height known

VoLoc needs to solve the following three sub-problems:

- Precisely estimate AoA for two signal paths in multipath-rich environments.
- Estimate the distance and orientation of at least one reflector, and identify the corresponding AoA for reverse triangulation.
- Fuse the AoAs, reflector, and height to geometrically infer the user's indoor location.

The solution needs to be performed without any voice training, must complete in the order of seconds, and must handle clutter in the environment (such as various objects scattered on the same table as Alexa).

3 SYSTEM ARCHITECTURE

Figure 4 illustrates *VoLoc*'s overall architecture. When the user speaks a voice command, the *IAC (Iterative Align-and-Cancel) AoA* module takes the raw acoustic samples, identifies the "pause" moment, and extracts a few initial AoAs from the following signal. To translate AoAs into location, the *Fusion* module takes two initial AoAs and fuses them with three parameters: the distance and orientation $\langle d, \theta \rangle$ of the nearby wall reflector, and the user's height, h. Together, the AoA and geometric information over-determine the user's 2D location for robustness to errors. The two parameters are separately estimated by the *Joint Geometric Parameter Estimation* module, by using the ensemble of *IAC*-estimated AoAs from recently received voice commands. This is a one-time estimation during initialization, meaning *VoLoc* is operational within the first $n = 15$ voice commands.

We begin this section by describing our IAC (Iterative Align-and-Cancel) AoA algorithm, a general AoA technique that also applies to other applications.

3.1 IAC (Iterative Align-and-Cancel) AoA

The goal of the IAC AoA algorithm is to extract both angles-of-arrival and corresponding delays of a few early paths in the multipath environment. This is very different from existing AoA algorithms which perform only *alignment* to find AoAs; we perform both *alignment and cancellation*, starting with a clean signal at the initial pause moment.

■ **A Glance at the Initial Moment**

Figure 5 zooms into the scenario when the voice signal is beginning to arrive at the microphones. The user starts to speak a sequence of voice samples, denoted as $x(t)$ = "ABCDE...". The signal travels along the direct (red) path, and arrives at the microphone array as early as time t_1. Note that due to the microphone arrangement in the array, mic #1, #2, $\cdots$ hear the first sample "A" at slightly different times: $t_1^{(1)}, t_1^{(2)}, \cdots$. These slight differences capture the AoA of the direct path.

With ensuing time, the same voice signal also arrives along the second (blue) path, known as the first echo. Since this second (blue) path is longer, the signal arrives at the microphone array at a later time, t_2, denoted as "abcdefg...". As a result, between t_1 and t_2, all the microphones hear clean, unpolluted direct path signal (tens of samples). Similarly, if t_3 is the time the third path arrives, then for $t \in [t_2, t_3]$, the microphones receive the signal from only the first two paths.

[2] The time granularity of a human voice command.

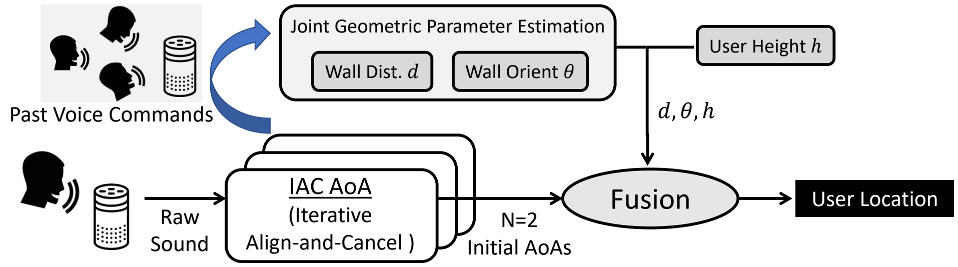


Figure 4: *VoLoc* system overview. When a user speaks a voice command, IAC *AoA* computes two initial AoAs. The direct path's AoA, when combined with height of the user, is ready to produce a basic 2D user location. To improve the estimate, the *Fusion* module fuses the two AoAs, the closest wall reflector, and the height information together to geometrically refine the location. The *Joint Parameter Estimation* module aims at computing the wall's relative distance and orientation, by analyzing recent voice commands from the user.

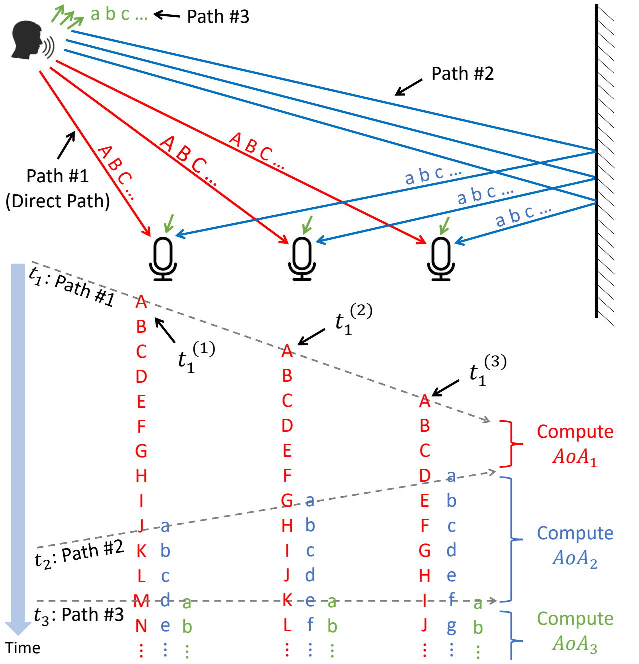


Figure 5: In a multipath environment, the voice signal travels along different paths and arrives at the microphone at different times.

■ **Detecting AoA_1 for the Direct Path**

Recall that signals in time window $t \in [t_1, t_2]$ contain only the direct path signal, and its angle-of-arrival (denoted as AoA_1) is captured in the slight time offset across microphones: $t_1^{(1)}, t_1^{(2)}, \cdots$. To infer AoA_1, we first detect t_1 from the rise of signal energy, and select a small time window $[t_1, t_1 + \Delta]$ of signals after that. Then, we ask the following question: Given this window of data, among all possible AoAs, which AoA_1 best aligns with the actual time offsets across the three microphones?

We solve this problem by performing a one-step "align-and-cancel". Figure 6(a) shows the key idea. Assume we have found the correct AoA_1; then, for any given pair of microphones, we can align their signals based on this AoA, in order to "reverse" the offset effect. This alignment is done by simply applying a cross-delay, i.e., delaying microphone i's signal with j's delay, and j's signal with i's delay.[3] The aligned signals are now subtracted, meaning they should now fully cancel each other with zero cancellation error. Any cancellation residue that we observe quantifies the error in the alignment, which further indicates the error in the AoA estimation. After searching across all possible AoAs, we choose the one which minimizes the sum of cancellation errors across all microphone pairs.

■ **Detecting AoA_2 for the Second Path**

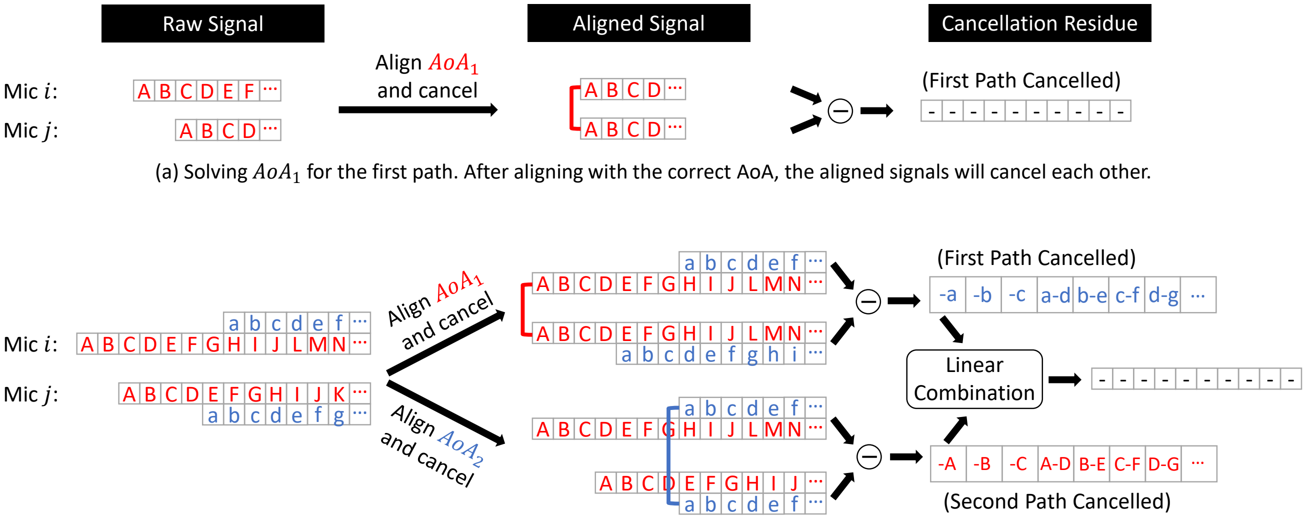
Once we have found the right AoA_1, the align-and-cancel operation should maintain low error over time, until when the second path arrives at a later time t_2. Thus, once we observe the error growing, it is time to estimate the second path's angle-of-arrival, AoA_2.

For this, we will again perform align-and-cancel for both AoA_1 and AoA_2, as shown in Figure 6(b). However, since the microphones are now receiving a mixture of two paths, we can align only one path at a time, meaning only one path gets canceled. In other words, after aligning the AoA_1 path and canceling the signals, the residue will be the difference between the "unaligned" second paths, and the vice versa. The middle column in Figure 6(b) shows both the alignments.

Fortunately, as shown in the third column of Figure 6(b), *the two cancellation residues are identical, except for a scaling factor* caused by the amplitude difference between two paths. This similarity is because both residues are essentially the "unaligned" path signal minus a delayed copy of the same "unaligned" signal, and that delay (which is the delay caused by AoA1 plus the delay caused by AoA2) is the same for both alignments. A linear combination of the two residues will be zero, and the coefficients are exactly the amplitudes of each path.

Based on the observation above, we solve for the second path's AoA by doing the following: We search across all possible AoAs for path #2, and for each AoA candidate, we perform the operation in

[3] It's easy to understand this by imagining the microphones' delays as two numbers x and y. To align them, we just need to add x to y and y to x, making both microphone's delays $(x + y)$.



(a) Solving AoA_1 for the first path. After aligning with the correct AoA, the aligned signals will cancel each other.

(b) Solving AoA_2 for the second path. The residual signals after "align and cancel" can further cancel each other by aligning the relative shift and scale.

Figure 6: The idea of iterative delay-and-cancellation (*IAC*) algorithm, shown for $K = 1$ and $K = 2$.

Figure 6(b), and run least squares (LS) over the two residues. The LS solution gives the estimated linear coefficients (which are the amplitudes), and the *fitting error* of LS indicates the cancellation error after alignment. We pick the candidate which has the smallest sum of fitting errors.

One point worth noting is that AoA only captures relative time offsets among microphones. To fully cancel the two residues, we also need to make sure the two cancellation residues (from first path and second path) are aligned in absolute time scale. This means the absolute delay t_2 has to be accurate as well. Since our t_2 detection may not be precise, we also jointly search for the correct t_2 around the time when the first path's cancellation error starts growing. We jointly pick the t_2 and AoA_2 that minimize the fitting error.

■ Detecting More AoAs

The same idea applies to getting more AoAs. If the received signal contains K paths, and assume we have obtained the AoAs (and absolute delays) for the first $(K-1)$ paths correctly, then we can search for the AoA (and absolute delay) of the K-th path by following operations similar to those in Figure 6(b). Theorem 3.1 states that when all K AoAs are estimated correctly, the K cancellation residues are linearly dependent, i.e., a linear combination of them (with coefficients as each path's amplitude) will be a zero vector. Therefore, searching for the K-th AoA is same as finding parameters that make K cancellation residues linearly dependent.

Theorem 3.1 (IAC AoA Decoding). *For any given pair of microphones, the K residue vectors – from aligning and canceling each of the K AoAs – are linearly dependent.*

Proof. Denote the source signal "ABCDEFG..." as $x[t]$, and the signal arriving along the k-th path at the i-th microphone as $x[t - t_{k,i}]$ (ignoring amplitude). Then, the total signal from all the K paths received by the i-th microphone can be written as: $y_i[t] = \sum_{k=1}^{K} x[t - t_{k,i}]$.

Now, when we align the k'-th path's AoA, the aligned signals at the two microphones become $y_1[t - t_{k',2}]$ and $y_2[t - t_{k',1}]$, respectively. The cancellation residue is:

$$\begin{aligned} & y_1[t - t_{k',2}] - y_2[t - t_{k',1}] \\ = & \sum_{k=1}^{K} x[t - t_{k,1} - t_{k',2}] - \sum_{k=1}^{K} x[t - t_{k,2} - t_{k',1}] \end{aligned}$$

The sum of all the cancellation residues (for all the K paths) is:

$$\begin{aligned} & \sum_{k'=1}^{K} \left(y_1[t - t_{k',2}] - y_2[t - t_{k',1}]\right) \\ = & \sum_{k'=1}^{K} \left(\sum_{k=1}^{K} x[t - t_{k,1} - t_{k',2}]\right) - \sum_{k'=1}^{K} \left(\sum_{k=1}^{K} x[t - t_{k,2} - t_{k',1}]\right) \\ = & \sum_{k=1}^{K} \sum_{k'=1}^{K} x[t - t_{k',1} - t_{k,2}] - \sum_{k'=1}^{K} \sum_{k=1}^{K} x[t - t_{k,2} - t_{k',1}] = 0 \end{aligned}$$

This proves that for any two microphones, the sum of cancellation residues is zero. Of course, here we have deliberately ignored amplitude (by assuming equal amplitude of each path). It is easy to prove that with the correct amplitude, the sum of cancellation vectors is still zero. Therefore, these residues are linearly dependent. □

Explained differently, observe that we obtain K residues after aligning-and-canceling each path. These K residues are linearly dependent (i.e., their sum is zero with correct coefficients). However, the linear coefficients are not known yet since the amplitudes are unknown to us. Therefore, to find the correct AoA_k, we simply search for different t_k and AoA_k, and run least squares fitting on the cancellation residues in order to minimize the fitting error. The best parameter is the one that achieves the minimal cancellation/fitting error. Algorithm 1 shows in detail how we compute this error.

Algorithm 1 For a Given Set of K AoAs and Absolute Delays, Compute the Overall Cancellation Error $\mathcal{E}$

```
Initialize overall error $\mathcal{E} = 0$
for all pairs of microphones do
    $ResidueVectors = \{\}$
    for each path $k = 1 \cdots K$ do
        Align the two signals using the $k$-th AoA
        Compute the difference of two aligned signals as the cancellation residue vector
        Delay the residue vector using the $k$-th absolute delay
        $ResidueVectors$.Add(residue)
    end for
    Run least squares on $ResidueVectors$ to compute the best linear combination, and get its fitting error $e$
    $\mathcal{E} = \mathcal{E} + e$
end for
```

■ Can We Detect Infinite AoAs?

In practice, the number of AoAs we could obtain is limited for two reasons: (1) In multipath environments, the first few paths are sparse (in time) while the latter ones are dense. This means the time window $[t_k, t_{k+1}]$ will be very short as k grows larger, making it hard to find the k-th path's AoA without being influenced by the $(k+1)$-th path. Said differently, there is no strict time of arrival of a given echo, hence, shorter gaps between arriving echoes make them difficult to separate. (2) Voice energy ramps up slowly due to the way humans produce sound. This means the latter echoes of the early samples are considerably weaker than the direct path samples. Background noise adds to this, further lowering the SNR of the latter echoes. This is why *VoLoc* conservatively uses only the first $N = 2$ AoAs.

■ Simulation Results

To compare *IAC*'s AoA estimation accuracy with other AoA algorithms under different indoor reverberation and SNR conditions, as well as to obtain ground truth for higher-order AoAs, we run a simulation with the "Alexa" voice as the source signal, added with varying levels of echoes and noise. The simulation uses the image source model [14] to simulate room impulse responses. We compare with three AoA techniques discussed in Section 2: (1) delay-and-sum, (2) MUSIC, and (3) GCC-PHAT.

Figure 7 shows the accuracy performance of these four algorithms. In general, we observe that GCC-PHAT is robust to reverberation and can get the first path correctly, but the correlation will fail at higher order paths. MUSIC and delay-and-sum do not work well in indoor reverberated environments where the acoustic signals are highly correlated. IAC, in contrary, actively takes advantage of correlated signals to jointly estimate each path's AoA and delay, leading to improved performance in rooms. This is the reason why our algorithm is, we believe, a new contribution to the body of AoA algorithms.

3.2 User Localization via Fusion

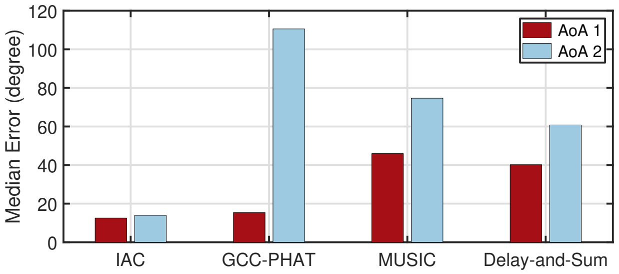


Figure 7: Accuracy comparison of four AoA techniques: IAC, GCC-PHAT, MUSIC, and Delay-and-Sum.

The above estimated AoAs can be reverse-triangulated to the user's location when we already know where the nearby wall reflector is, i.e., its distance d and orientation θ with respect to Alexa. Moreover, the human height (h) constrains the location search space to 2D. Pretending we already know $\langle d, \theta, h\rangle$, we design an optimization technique to efficiently fuse all three parameters to infer user location. In the next section, we will discuss how we jointly infer the $\langle d, \theta\rangle$ parameters from recent voice signals.

In ideal settings, the two AoAs and the wall's $\langle d, \theta\rangle$ are enough to analytically solve for the source location. In real environments, all the AoA and geometry estimates incur error, so over-determining the system with echo delay and human height h is valuable. In fusing all these and solving for user location, we make the following observations and decisions:

(1) First, not all AoAs/delays are feasible as the user is only moving in 2D with a fixed height. Therefore, searching for user location in this 2D plane will be more efficient (than searching for all AoAs and delays).

(2) Second, the direct path AoA from *IAC*, especially its azimuth, is accurate. This further reduces the search space to a beam in the 2D plane, as shown in Figure 8.

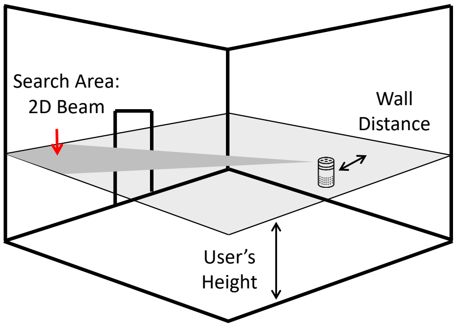


Figure 8: The reduced search space.

(3) Finally, for each possible location on this 2D beam, we can directly obtain the parameters for the direct path and wall reflection path using geometry (recall, we pretended to know all three parameters). This means we can directly compute the cancellation error using Algorithm 1 (using $K = 2$ echoes). The best location is determined by having the minimum cancellation error. Algorithm 2 summarizes our searching procedure. While this procedure is a highly non-convex optimization, the search finishes within a few seconds due to limited search space.

Algorithm 2 Search for the Most Likely User Location in the Room

```
Run $IAC$ to first obtain direct path's azimuth, $azi$
$minError = +\infty$,    $bestLoc = [\ ]$
**for all** Location $loc$ on 2D plane **do**
    **if** $loc$'s azimuth not close to $azi$ **then**
        **continue**
    **end if**
    Compute AoA and absolute delay for both direct path and
    wall reflection path, using geometry
    Compute cancellation error $\mathcal{E}$ using Algorithm 1
    **if** $\mathcal{E} < minError$ **then**
        $minError = \mathcal{E}$,    $bestLoc = loc$
    **end if**
**end for**
Declare $bestLoc$ as user location
```

3.3 Joint Wall Parameter Estimation

Finally, we describe our solution to estimate the two parameters (d, θ) from an ensemble of recent voice samples. Our core idea is the following. For each (past) voice command, we utilize the direct path AoA as a trustworthy estimate. We shortlist locations within a 3D cone around this AoA that satisfies our known height h. Now for each of these locations, and pretending the wall is d_i, θ_j from Alexa, we compute the corresponding wall AoA and delay. If $\langle d_i, \theta_j \rangle$ are the correct estimates, then the computed AoA and delay will align well with the measured signal, minimizing the cancellation error in Algorithm 1.

Of course, in the presence of other echoes from clutters around Alexa, the wall echo may not match best, hence $\langle d_i, \theta_j \rangle$ may produce a higher (than minimum) cancellation error. However, when this operation is performed over multiple recent voice commands, and the cancellation errors summed up, we expect the optimal $\langle d_i^*, \theta_j^* \rangle$ to minimize this sum. The intuition is that different voice commands from different locations would consistently reflect from the wall, while reflections from other objects would come and go.[4] As a result, the correct values of $\langle d_i^*, \theta_j^* \rangle$ would eventually "stand out" over time.

Figure 9 shows one example of how the objective function (i.e., sum of cancellation errors) varies across the joint $\langle d, \theta \rangle$ variation. The X and Y axes of the graph are d and θ offsets from the ground truth, meaning the contour should minimize at $[0, 0]$. We search with a granularity of 2 cm and 1°, and the minimum point of the contour proves to be at $X = 2$ cm and $Y = 1°$. This is promising and we evaluate this more in Section 4.

While joint parameter estimation is time consuming (in hours), we need to run this only during initialization. Once the estimates are ready, the fusion module uses these parameters and executes in a very short time.

3.4 Points of Discussion

■ **Will echoes from furniture / environment affect the estimation of wall geometry?**

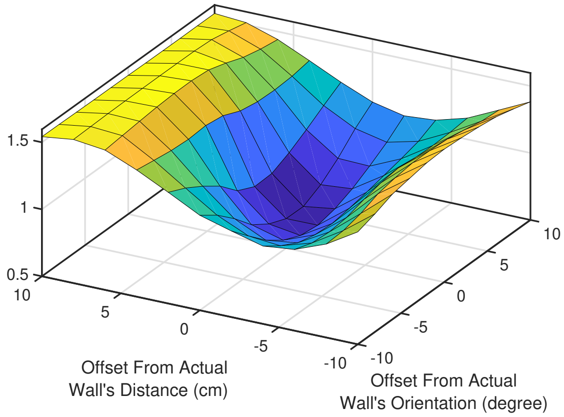


Figure 9: The sum of cancellation error minimizes at parameters that are close to the actual distance and orientation.

Observe that the echo from the nearby wall is also close in time to the direct path. In fact, the echo's delay can be computed since $\langle d^*, \theta^*, h \rangle$ are all known. Because of this, echoes that bounce off farther away reflectors can be discounted, since all their delays arrive long after the wall-echo. Confusion arises from reflectors that are closer to Alexa than the wall – like objects on the same table as Alexa. These un-modeled echoes prevent the cancellation errors from dropping sharply. Nonetheless, as we see in our evaluation, the error reduction is still a minimum for the correct user location. This is the reason why *VoLoc* is able to operate even in reasonably cluttered environments.

■ **What happens when the wall echo is blocked by an object on the table?**

This is the case where *VoLoc* will perform poorly, since the cancellation will be poor for the expected wall AoA. It may be possible to recognize the problem from the value of the cancellation error, such that we can gain some measure of confidence on the localization result. We have empirically observed increased cancellation errors; however, it is not clear how to develop a systematic confidence score from it (note that global thresholds or distributions would not scale well). Hence, we leave the design of a confidence metric to future work.

4 IMPLEMENTATION, EVALUATION

This section discusses the experiment methodology and performance results of *VoLoc*.

4.1 Implementation

VoLoc is implemented on a Seeed Studio 6-microphone array [11], arranged in a circular shape similar to Amazon Echo. This is due to the lack of raw acoustic samples from commercial voice assistants. The acoustic signals on the array are sampled at 16 kHz, a sampling rate that covers most of the energy in voice frequencies, and further high-pass filtered at 100 Hz to eliminate background noise. The array is connected to a Raspberry Pi to forward its sound samples to a laptop over wireless. The laptop executes the code written in MATLAB to compute user location, which takes 6 – 10 seconds to finish.

[4] The table reflection may also be consistent; however, Alexa/Google microphones are designed with low gain towards the downward direction, and hence the energy of table reflection is weak.

4.2 Methodology

Our experiments were executed in four different indoor environments: (1) a studio apartment, (2) a kitchen, (3) a student office, and (4) a large conference room. The first two in-home places both have an Amazon Echo pre-installed, so we directly replace it with our microphone array. For the office and the conference room, we simply put the microphone array on a desk that is close to a power outlet. The distance to the close-by wall ranges between 0.2 m and 0.8 m.

We recruited three student volunteers to speak different voice commands to the microphone array. Volunteers were asked to stand at marked positions, whose 2D locations (X, Y) have been measured beforehand (for ground truth) using a laser distance measurer. The voice commands start with either "Alexa, ..." or "Okay Google, ...", and are repeated five times at each location. We collected a total number of 2350 voice commands. Meanwhile, for in-home environments, we also recorded some other non-voice sounds and played them at different locations using a portable Bluetooth speaker. These sounds include the sound of cooking, the microwaves dings, or random sound clips from TVs. The goal is to test whether *VoLoc* has the potential to localize such arbitrary sounds from everyday objects.

4.3 Performance Results

The following questions are of interest:

(1) How well can *VoLoc* compute user locations in general? What is the break-up of gain from AoA and wall-estimation?

(2) How does *VoLoc*'s performance vary among different sounds (including non-voice sounds), varying clutter level, and varying ranges (near, medium, far)?

(3) How many recent voice samples are necessary for *VoLoc* to converge on the geometric parameters (d, θ)?

■ **Overall Localization Accuracy**

Figure 10 shows the CDF of *VoLoc*'s overall localization errors across all experiments, as well as the CDF of errors in each room. Overall, the median error is 0.44 m. We believe this accuracy makes it amenable to location-aware applications for in-home voice assistants like Alexa and Google Home.

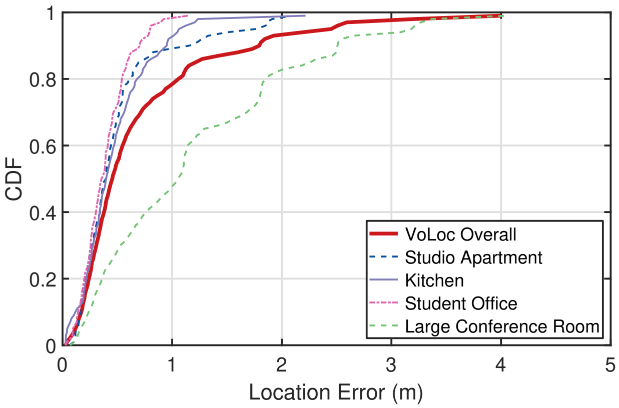


Figure 10: CDF of *VoLoc*'s overall localization accuracy, and the accuracy across different rooms.

Upon comparing the performance across the four rooms, we find that the conference room incurs significantly higher errors than the other three. Analysis shows that the conference room is large in size, meaning the user often stands far from Alexa, leading to increased location error. Said differently, far field errors are higher in triangulation algorithms because same angular error (in AoA, d, or θ) translates to larger location error.

Figure 11 compares *VoLoc*'s localization accuracy with the following two schemes: (1) *VoLoc*++, which assumes the two geometric parameters (wall's distance d and orientation θ) are perfectly known. Therefore, *VoLoc*++ will be a performance upper bound of *VoLoc*. (2) GCC-PHAT, which combines GCC-PHAT's direct path AoA with human height information (h) to compute human location. We choose GCC-PHAT as the baseline because it performs the best in Section 3.1.

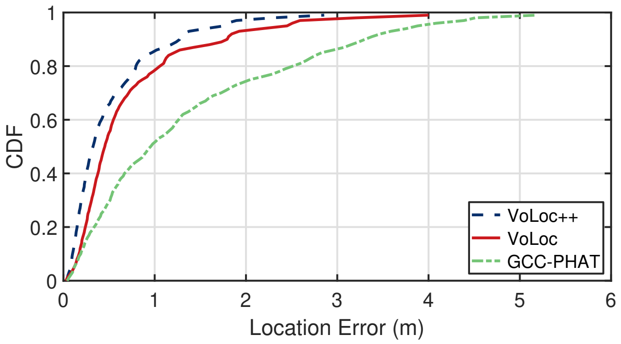


Figure 11: Performance comparison of *VoLoc*++, *VoLoc*, and GCC-PHAT.

Compared to GCC-PHAT's median location error of 0.93 m, *VoLoc*'s median error reduces by 52%. This demonstrates the value of precise 2-AoA estimation from our *IAC* algorithm. *VoLoc*++ further reduces the median error from 0.44 m to 0.31 m, assuming the geometric parameters are precisely known. This captures *VoLoc*'s efficacy to estimate the wall parameters – there is a small room for improvement.

■ **Accuracy Across Different Sounds**

Figure 12 shows *VoLoc*'s median localization error across various kinds of sounds for in-home environments. The first two types of sounds are human voice commands, while the latter four are natural sounds from everyday objects, such as microwave bell sound or music from TV. In general, we observe that localizing objects' sounds is easier than localizing the human voice. This is because most sounds made by objects have good energy ramp-up property; i.e., unlike human voice, the energy of the sound quickly goes up within a very short time window. This means the SNR of the signal is strong for *IAC* to estimate AoA, leading to improved location results.

■ **Accuracy Over Distances to Alexa**

VoLoc's localization accuracy will naturally go down as the user moves away from the microphone array. This is essentially because the resolution of AoA estimation limits the range of the user, i.e., a large movement at a far away distance may only translate to a slight change in AoA. Figure 13 visualizes *VoLoc*'s localization error

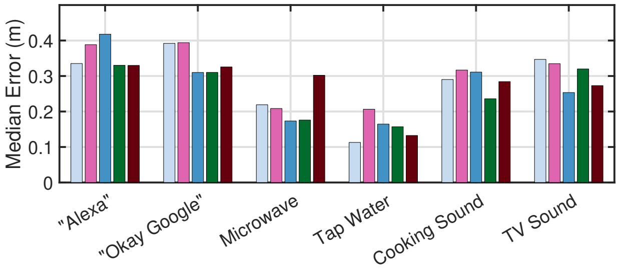


Figure 12: *VoLoc*'s localization accuracy across different kinds of sounds. Each cluster of bars represents one sound, and each bar within one cluster represents the median error across locations during one session.

across different locations in the conference room. The microphone array is placed on a table towards the northeast side. Evidently, the location accuracy varies with the proximity to the microphone array.

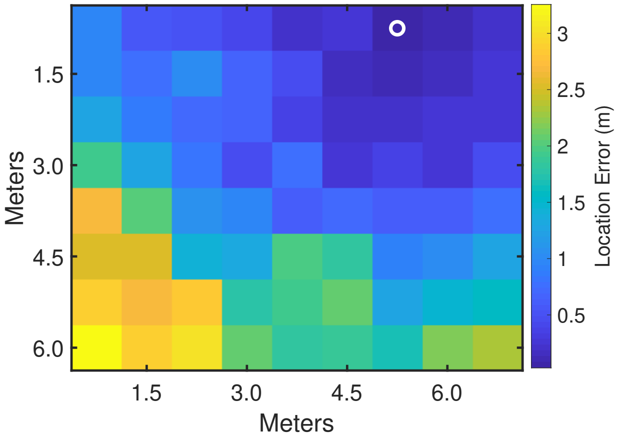


Figure 13: Heatmap of *VoLoc*'s localization error in the conference room (bird-eye view). Small white circle represents the microphone array location.

We classify all our measurements into three groups, based on the distance from the user to the microphone array: Near (< 2 m), Medium (2 – 4 m), and Far (> 4 m). Figure 14 shows the location accuracy for each group. We observe that within 2 m, location error is almost always below 0.5 m, and within 4 m, the majority of the errors are still within 1 m.

■ **Sensitivity to Different Users**

To test *VoLoc*'s sensitivity to different users, we asked three volunteers to enter the same room, stand at the same set of locations, and speak the same voice commands. Figure 15 shows the variation of median localization error across different locations. Evidently, localization error is much more correlated with the user's standing location (as would be expected), rather than the users voice or speaking patterns.

■ **Sensitivity to the Clutter Levels**

Clearly, *VoLoc*'s performance will depend on the density of the multipath signals due to other objects' reflections. Since we only look at the very beginning moment of the sound, most indoor reflections (like those from furniture) are not a problem for us. However, objects that are very close to the microphone array may reflect sounds into the microphone even earlier than the wall, or even totally block the wall reflection, leading to higher cancellation residue and more location errors. In the extreme case where even the direct path is totally blocked, the localization error will go up dramatically.

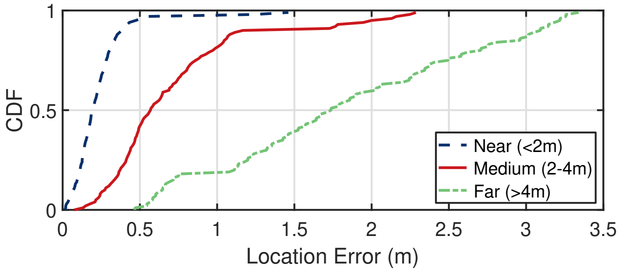


Figure 14: CDF of *VoLoc*'s localization error, for three different distance categories of the user from Alexa: near, medium, and far.

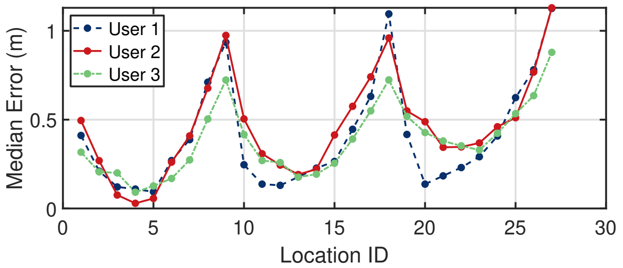


Figure 15: Variation of *VoLoc*'s localization error across different locations in the room, shown separately for each user. Location ID is labeled in a row-by-row manner.

Figure 16 shows the degradation of *VoLoc*'s localization accuracy, as we keep increasing the clutter level around the microphone array (i.e., putting objects on the same table as the array to add complexity to its multipath profile). Evidently, the error is low when there is no object nearby. Even when there are a few objects around and the clutter level is moderate, the location error is still acceptable. However, as more and larger objects start to fully block the wall reflection and even the direct path, the location error quickly increases.

■ **Sensitivity to Background Noise**

Our experiments were performed in reasonably quiet rooms, with ambient noises from home appliances such as refrigerators and computers. With increased background noise, *VoLoc* will not be able to achieve zero *IAC* cancellation residue, leading to increased errors in sound source localization.

Figure 17 shows how *VoLoc*'s performance degrades, as we take one set of microphone recordings inside a quiet room, manually add synthesized noises to the recordings, and then re-run the algorithm to localize the sound source. We test with two types of noises: (1) low-frequency machine humming noise, and (2) all-frequency

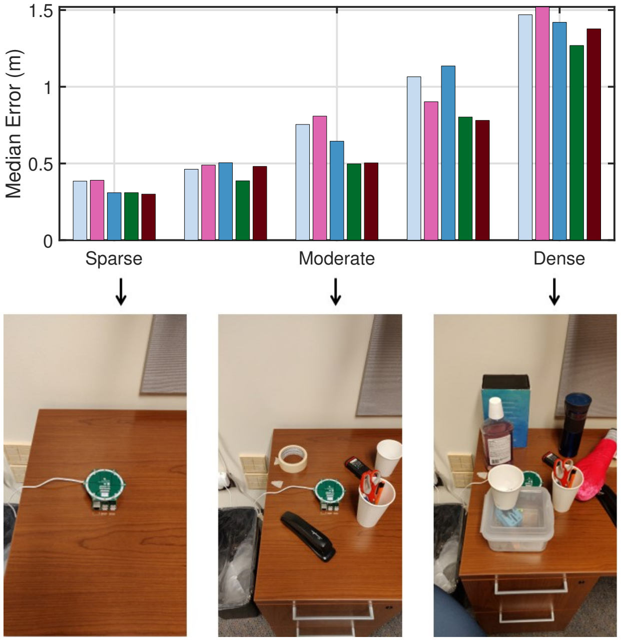


Figure 16: *VoLoc*'s localization accuracy across increasing clutter levels (from left to right). Each cluster of bars represents one environment (one clutter level), and each bar within one cluster represents the overall median error across different locations in the room during one visit. The three pictures correspond to the measurement in the #1, #3 and #5 environments.

background music. We observe slight performance degradation with humming noise, essentially because *VoLoc* employs a high-pass filter to eliminate low-frequency ambient noise and extract human voice signals. In the case of background music, the performance drops significantly. This is because music shares similar frequency bands to that of human voice, which makes the separation difficult. In the future, we plan to employ more fine-grained filters (instead of a simple high-pass filter) to further separate human voice from the recordings.

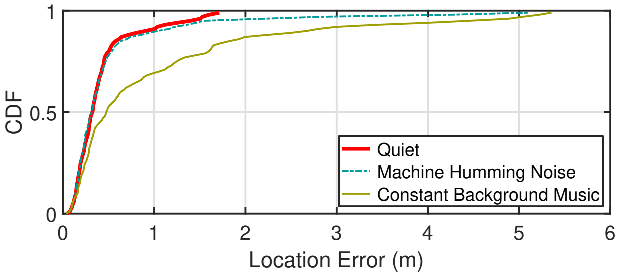


Figure 17: How *VoLoc*'s accuracy decreases with different noise characteristics. Noise reduces the SNR at the initial moment of the voice signal, leading to increased location error.

■ Convergence of Geometric Parameter Estimation

Figure 18 shows how the wall parameter estimation is converging, with an increasing number of past voice commands. While more past samples are useful, with as few as 5 samples, our estimation has converged to < 1 cm and $< 2°$ fluctuation, for the wall's distance and orientation, respectively. This shows *VoLoc*'s ability to converge to new wall parameters with a few voice samples, even after being moved around on the table. This experiment was performed at the medium clutter level (as per expectations, the estimation converges faster and slower for sparse and dense clutter levels, respectively).

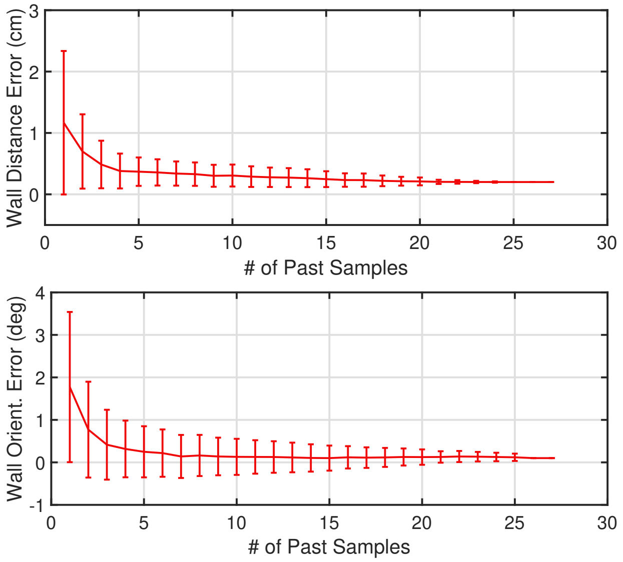


Figure 18: How *VoLoc*'s parameter estimation converges for (1) wall distance d, and (2) wall orientation θ, with increasing number of past voice samples. The red line and its error bar represent the average and standard deviation of estimation errors in distance and orientation, for a total number of 1000 runs.

5 LIMITATIONS AND DISCUSSION

In this section, we discuss limitations of *VoLoc* and room for improvement.

■ Semantic interpretation of location: *VoLoc* infers the user location and wall parameters in Alexa's reference frame. To be semantically meaningful (i.e., the user is at the laundry room), the inferred locations need to be superimposed on a floorplan [47, 48, 57]. Alternatively, Alexa could localize other sounds, understand their semantics, and transfer those semantics to location. For instance, knowing that the washer/dryer sound arrives from the same location as a voice command can reveal that the user is at the laundry room. Building such a semantic layer atop localization is an important follow-up work.

■ Line-of-sight path: *VoLoc* assumes that the line-of-sight path (direct path) exists between the user and Alexa. If the direct path is blocked (e.g., the user and Alexa are in different rooms), even the first path will contain reflections, and reverse triangulation may converge to a very different direction. Correcting this result requires an accurate knowledge of the indoor reflections, and we leave this to future work.

■ **Coping with variations in height:** A family will likely have multiple users with different heights (including children). *VoLoc* needs the height of each user and some form of voice fingerprinting to apply the corresponding height during computation. We have not implemented such per-user adaptation. We also do not cope with scenarios where the user is sitting or lying down (we assume standing users).

■ **Mobile users:** *VoLoc* has been designed and evaluated with static users. When a user issues a voice command while walking, the echo patterns will likely "spread" in space. Our current formulation does not model the effects of mobility – the algorithms will need to be revisited.

■ **Many pause opportunities:** A voice command offers at least two pause opportunities, one before the command, and one after the word "Alexa". Naively averaging location, derived from each pause, will improve accuracy. However, averaging in the signal space (i.e., optimizing the wall parameters using all the post-pause signals) could offer greater benefits. We leave such refinements to future work.

■ **Privacy:** How applications use the location information in the future remains an open question. On one hand, we see context-aware Alexas and Google Homes becoming crucial supporting technologies to old-age independent living; sharing the user's height may be worthwhile in such cases. On the other hand, for everyday users, we see Amazon and Google peering even more closely into our homes and daily lives, a stronger erosion of privacy. Overall, we believe indoor location might be an incremental privacy consideration, given that the user's voice is already being recorded and its AoA is already being computed (and actively shown to the user). Regardless of utility or abuse, we believe awareness of such capabilities is critical, and we believe this work is still important to spread awareness of what is possible from voice signals.

6 RELATED WORK

In the interest of space, we give a heavily condensed summary of the vast literature in sound source localization and acoustic signal processing, with bias towards the work that are more related to *VoLoc*.

■ **Multiple arrays or known sound signals**: Distributed microphone arrays have been used to localize (or triangulate) an unknown sound source, such as gun shots [63], wildlife [24], noise sources [55, 70, 71], and mobile devices [25]. Many works also address the inverse problem of localizing microphones with speaker arrays that are playing known sounds [13, 15, 42, 50]. Ishi et al. [37] report modeling the room multipath to improve multi-array localization results. Xiong and Jamieson [78] and Michalevsky et al. [49] demonstrate localization using multiple RF-based landmarks. On the other hand, when the source signal is known, localization has been accomplished by estimating the channel impulse response (CIR). For instance, [28] uses an acoustic sine sweep to localize room boundaries and compute the shape of a room; reverberations captured by multiple microphones reveal the room impulse responses (RIR), which stipulate the locations of reflecting surfaces. In RF (like WiFi), CIR and SNR based fingerprinting has been used extensively [17, 19, 53, 66, 79, 80]. As mentioned earlier, *VoLoc* must cope with a single array and unknown signals.

■ **Unknown signal, single array**: Perhaps closest to *VoLoc* are [16, 59]. In [16], a robot first scans the 3D model of an empty room with a Kinect depth sensor. It then identifies multipath AoAs of a clapping sound with a microphone array, and performs 3D reverse ray-tracing to localize the sound source position. In acoustics, the clapping sound is known as an impulse sound [10, 56, 65], making it trivial to estimate the channels and separate the echoes at the microphones. Ribeiro et al. [59] localize a sound source with a microphone array using the maximum likelihood approach. They train the reflection coefficients in an empty room, and present results from three carefully chosen locations to control the multipaths. The speaker-microphone distances are only about 1 m, essentially making it a near-field localization problem. In comparison, our solution is designed for real-world, multipath-rich, uncontrolled environments. In [33], a single microphone is used to classify a speaker's distance into a set of discrete values, but needs per-room, per-distance training.

■ **AoA estimation**: Rich bodies of work have focused on estimating acoustic AoAs using microphone arrays [21, 51, 52, 54, 73]. Some are best suited for different sound sources, some for specific signal types and frequencies, some for certain environments. Examples include delay-and-sum [30, 58, 77], GCC-AoA [18, 20, 32, 39, 76], MUSIC [64, 72, 74], and ESPRIT [61, 67]. However, in multipath-rich environments, blind AoA estimation struggles, especially for successive AoAs.

■ **Blind channel estimation**: Blind channel estimation (BCE) describes the process of deducing a channel from received signals without knowing the source signal. BCE is a useful tool for estimating indoor acoustic channels [26, 40, 44, 45]. We consider IAC to be a particular and powerful realization of BCE with significant computation gains. IAC was also inspired by ZigZag decoding for RF networks [34], which decodes packets by exploiting interference-free segments.

7 CONCLUSION

This paper shows the feasibility of inferring user location from voice signals received over a microphone array. While the general problem is extremely difficult, we observe that application-specific opportunities offer hope. Instead of inferring and triangulating all signal paths in an indoor environment, we observe that estimating a few AoAs and reflector surfaces is adequate for the localization application. We design an iterative cancellation algorithm (IAC) for AoA estimation, followed by a joint optimization of wall distance and orientation. When fused together, the localization accuracies are robust and usable.

ACKNOWLEDGMENTS

We sincerely thank the anonymous shepherd and reviewers for their insightful comments and suggestions. We are also grateful to Samsung, Intel, Google, Qualcomm, HP, and NSF (grant CNS-1719337 and CNS-1910933) for partially funding this research.

REFERENCES

[1] 2018. Audio Hardware Configurations. Retrieved Oct 30, 2019 from https://developer.amazon.com/docs/alexa-voice-service/audio-hardware-configurations.html

[2] 2018. How do you organize your devices? Naming conventions? (2018). Retrieved Oct 30, 2019 from https://community.smartthings.com/t/how-do-you-organize-your-devices-naming-conventions-2018/121688

[3] 2018. Qualcomm delivers a fully integrated far-field smart audio reference platform for the Amazon Alexa voice service. Retrieved Oct 30, 2019 from https://www.qualcomm.com/news/releases/2018/01/08/qualcomm-delivers-fully-integrated-far-field-smart-audio-reference-platform

[4] 2018. Siri gets smarter at recognizing names of local businesses and attractions. Retrieved Oct 30, 2019 from https://voicebot.ai/2018/08/13/siri-gets-smarter-at-recognizing-names-of-local-businesses-and-attractions/

[5] 2019. Alexa, lights! How I turned my home into a sci-fi dream. Retrieved Oct 30, 2019 from https://www.theguardian.com/technology/2016/dec/23/alexa-lights-how-i-turned-my-home-into-a-sci-fi-dream

[6] 2019. Amazon Files for Patent to Detect User Illness and Emotional State by Analyzing Voice Data. Retrieved Oct 30, 2019 from https://voicebot.ai/2018/10/10/amazon-files-for-patent-to-detect-illness-by-analyzing-voice-data/

[7] 2019. Audio Analytic: Enabling intelligent products through sound recognition. Retrieved Oct 30, 2019 from https://www.audioanalytic.com/

[8] 2019. Beco Focuses on Developing a Spatially-Aware Alexa Skill. Retrieved Oct 30, 2019 from https://developer.amazon.com/blogs/alexa/post/Tx1BPHXBLZV5ZVN/beco-focuses-on-developing-a-spatially-aware-alexa-skill

[9] 2019. Here are the upcoming Alexa features we're most excited about. Retrieved Oct 30, 2019 from https://thenextweb.com/artificial-intelligence/2018/09/20/here-are-the-upcoming-alexa-features-were-most-excited-about/

[10] 2019. Impulse Responses, Odeon A/S. Retrieved Oct 30, 2019 from https://odeon.dk/downloads/impulse-responses/

[11] 2019. ReSpeaker 6-mic circular array kit for Raspberry Pi. Retrieved Oct 30, 2019 from http://wiki.seeedstudio.com/ReSpeaker_6-Mic_Circular_Array_kit_for_Raspberry_Pi/

[12] 2019. 'Take your pill!' Ultra-precise Dot beacons trigger alerts that matter. Retrieved Oct 30, 2019 from https://finance.yahoo.com/news/pill-ultra-precise-dot-beacons-181110596.html

[13] Teodoro Aguilera, José A Paredes, Fernando J Álvarez, José I Suárez, and A Hernandez. 2013. Acoustic local positioning system using an iOS device. In *International Conference on Indoor Positioning and Indoor Navigation*. IEEE, 1–8.

[14] 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, 4 (1979), 943–950.

[15] Fernando J Álvarez, Teodoro Aguilera, and Roberto López-Valcarce. 2017. CDMA-based acoustic local positioning system for portable devices with multipath cancellation. *Digital Signal Processing* 62 (2017), 38–51.

[16] Inkyu An, Myungbae Son, Dinesh Manocha, and Sung-Eui Yoon. 2018. Reflection-Aware Sound Source Localization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 66–73.

[17] P Bahl and VN Padmanabhan. 2000. RADAR: An in-building RF-based user location and tracking system. In *Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No. 00CH37064)*, Vol. 2. IEEE, 775–784.

[18] Jacob Benesty, Jingdong Chen, and Yiteng Huang. 2004. Time-delay estimation via linear interpolation and cross correlation. *IEEE Transactions on Speech and Audio Processing* 12, 5 (2004), 509–519.

[19] Dinesh Bharadia, Emily McMilin, and Sachin Katti. 2013. Full duplex radios. In *ACM SIGCOMM Computer Communication Review*, Vol. 43. ACM, 375–386.

[20] Michael S Brandstein and Harvey F Silverman. 1997. A robust method for speech signal time-delay estimation in reverberant rooms. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, Vol. 1. IEEE, 375–378.

[21] Alessio Brutti, Maurizio Omologo, and Piergiorgio Svaizer. 2011. Inference of acoustic source directivity using environment awareness. In *2011 19th European Signal Processing Conference*. IEEE, 151–155.

[22] Daguan Chen. 2019. *Source localization from voice signals*. Master's thesis. University of Illinois at Urbana-Champaign.

[23] A. Chhetri, P. Hilmes, T. Kristjansson, W. Chu, M. Mansour, X. Li, and X. Zhang. 2018. Multichannel Audio Front-End for Far-Field Automatic Speech Recognition. In *2018 26th European Signal Processing Conference (EUSIPCO)*. 1527–1531. https://doi.org/10.23919/EUSIPCO.2018.8553149

[24] Travis C Collier, Alexander NG Kirschel, and Charles E Taylor. 2010. Acoustic localization of antbirds in a Mexican rainforest using a wireless sensor network. *The Journal of the Acoustical Society of America* 128, 1 (2010), 182–189.

[25] Ionut Constandache, Sharad Agarwal, Ivan Tashev, and Romit Roy Choudhury. 2014. Daredevil: Indoor location using sound. *ACM SIGMOBILE Mobile Computing and Communications Review* 18, 2 (2014), 9–19.

[26] Marco Crocco and Alessio Del Bue. 2015. Room impulse response estimation by iterative weighted L1-norm. In *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 1895–1899.

[27] Joseph Hector DiBiase. 2000. *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*. Brown University Providence, RI.

[28] Ivan Dokmanić, Reza Parhizkar, Andreas Walther, Yue M. Lu, and Martin Vetterli. 2013. Acoustic echoes reveal room shape. https://www.pnas.org/content/110/30/12186. *Proceedings of the National Academy of Sciences* 110, 30 (2013), 12186–12191. https://doi.org/10.1073/pnas.1221464110 arXiv:https://www.pnas.org/content/110/30/12186.full.pdf

[29] Ivan Dokmanic and M Vetterli. 2015. Listening to distances and hearing shapes: Inverse problems in room acoustics and beyond. *EPFL, Lausanne* (2015).

[30] Gary W Elko. 1996. Microphone array systems for hands-free telecommunication. *Speech Communication* 20, 3-4 (1996), 229–240.

[31] Mark Edward Epstein and Lucy Vasserman. 2016. Generating language models. US Patent 9,437,189.

[32] Izabela L Freire et al. 2011. DOA of gunshot signals in a spatial microphone array: Performance of the interpolated generalized cross-correlation method. In *2011 Argentine School of Micro-Nanoelectronics, Technology and Applications*. IEEE, 1–6.

[33] Eleftheria Georganti, Tobias May, Steven van de Par, Aki Harma, and John Mourjopoulos. 2011. Speaker distance detection using a single microphone. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 7 (2011), 1949–1961.

[34] Shyamnath Gollakota and Dina Katabi. 2008. *Zigzag Decoding: Combating Hidden Terminals in Wireless Networks*. Vol. 38. ACM.

[35] Larry Paul Heck, Madhusudan Chinthakunta, David Mitby, and Lisa Stifelman. 2016. Location based conversational understanding. US Patent 9,244,984.

[36] J. Heymann, M. Bacchiani, and T. N. Sainath. 2018. Performance of Mask Based Statistical Beamforming in a Smart Home Scenario. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6722–6726. https://doi.org/10.1109/ICASSP.2018.8462372

[37] Carlos Toshinori Ishi, Jani Even, and Norihiro Hagita. 2013. Using multiple microphone arrays and reflections for 3D localization of sound sources. *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems* (2013), 3937–3942.

[38] Huafeng Jin and Shuo Wang. 2018. Voice-based determination of physical and emotional characteristics of users. US Patent App. 15/457,846.

[39] Charles Knapp and Glifford Carter. 1976. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24, 4 (1976), 320–327.

[40] Konrad Kowalczyk, Emanuël AP Habets, Walter Kellermann, and Patrick A Naylor. 2013. Blind system identification using sparse learning for TDOA estimation of room reflections. *IEEE Signal Processing Letters* 20, 7 (2013), 653–656.

[41] Swarun Kumar, Stephanie Gil, Dina Katabi, and Daniela Rus. 2014. Accurate indoor localization with zero start-up cost. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*. ACM, 483–494.

[42] Patrick Lazik, Niranjini Rajagopal, Oliver Shih, Bruno Sinopoli, and Anthony Rowe. 2015. ALPS: The Acoustic Location Processing System. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*. ACM, 491–492.

[43] Xinyu Lei, Guan-Hua Tu, Alex X Liu, Chi-Yu Li, and Tian Xie. 2018. The insecurity of home digital voice assistants-vulnerabilities, attacks and countermeasures. In *2018 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 1–9.

[44] Yuanqing Lin, Jingdong Chen, Youngmoo Kim, and Daniel D Lee. 2007. Blind sparse-nonnegative (BSN) channel identification for acoustic time-difference-of-arrival estimation. In *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 106–109.

[45] Yuanqing Lin, Jingdong Chen, Youngmoo Kim, and Daniel D Lee. 2008. Blind channel identification for speech dereverberation using l1-norm sparse learning. In *Advances in Neural Information Processing Systems*. 921–928.

[46] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech Model Pre-training for End-to-End Spoken Language Understanding. *arXiv Preprint arXiv:1904.03670* (2019).

[47] Rufeng Meng, Sheng Shen, Romit Roy Choudhury, and Srihari Nelakuditi. 2016. Autolabel: Labeling places from pictures and websites. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1159–1169.

[48] Rufeng Meng, Sheng Shen, Romit Roy Choudhury, and Srihari Nelakuditi. 2015. Matching physical sites with web sites for semantic localization. In *Proceedings of the 2nd Workshop on Physical Analytics*. ACM, 31–36.

[49] Yan Michalevsky, Aaron Schulman, Gunaa Arumugam Veerapandian, Dan Boneh, and Gabi Nakibly. 2015. Powerspy: Location tracking using mobile device power analysis. In *24th USENIX Security Symposium (USENIX Security 15)*. 785–800.

[50] João Neves Moutinho, Rui Esteves Araújo, and Diamantino Freitas. 2016. Indoor localization with audible sound - Towards practical implementation. *Pervasive and Mobile Computing* 29 (2016), 1–16.

[51] Kazuhiro Nakadai, Hirofumi Nakajima, Kentaro Yamada, Yuji Hasegawa, Takahiro Nakamura, and Hiroshi Tsujino. 2005. Sound source tracking with directivity pattern estimation using a 64 ch microphone array. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1690–1696.

[52] Hirofumi Nakajima, Keiko Kikuchi, Toru Daigo, Yutaka Kaneda, Kazuhiro Nakadai, and Yuji Hasegawa. 2009. Real-time sound source orientation estimation using a 96 channel microphone array. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 676–683.

[53] Rajalakshmi Nandakumar, Krishna Kant Chintalapudi, Venkat Padmanabhan, and Ramarathnam Venkatesan. 2013. Dhwani: Secure peer-to-peer acoustic NFC. In *ACM SIGCOMM Computer Communication Review*, Vol. 43. ACM, 63–74.

[54] Kenta Niwa, Yusuke Hioka, Sumitaka Sakauchi, Ken'ichi Furuya, and Yoichi Haneda. 2010. Estimation of sound source orientation using eigenspace of spatial correlation matrix. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 129–132.

[55] Jean-Francois Piet, Georges Elias, Jean-Francois Piet, and Georges Elias. 1997. Airframe noise source localization using a microphone array. In *3rd AIAA/CEAS Aeroacoustics Conference*. 1643.

[56] Thomas Plotz and Gernot A Fink. 2008. On the use of empirically determined impulse responses for improving distant talking speech recognition. In *2008 Hands-Free Speech Communication and Microphone Arrays*. IEEE, 156–159.

[57] Niranjini Rajagopal, Patrick Lazik, and Anthony Rowe. 2014. Visual light landmarks for mobile devices. In *Proceedings of the 13th International Symposium on Information Processing in Sensor Networks*. IEEE, 249–260.

[58] António LL Ramos, Sverre Holm, Sigmund Gudvangen, and Ragnvald Otterlei. 2011. Delay-and-sum beamforming for direction of arrival estimation applied to gunshot acoustics. In *Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense X*, Vol. 8019. International Society for Optics and Photonics, 80190U.

[59] Flavio P. Ribeiro, Demba E. Ba, Cha Zhang, and Dinei A. F. Florêncio. 2010. Turning enemies into friends: Using reflections to improve sound source localization. *2010 IEEE International Conference on Multimedia and Expo* (2010), 731–736.

[60] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. 2018. Inaudible voice commands: The long-range attack and defense. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*. 547–560.

[61] Richard Roy and Thomas Kailath. 1989. ESPRIT-estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37, 7 (1989), 984–995.

[62] Tara N Sainath, Ron J Weiss, Kevin W Wilson, Arun Narayanan, Michiel Bacchiani, et al. 2015. Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 30–36.

[63] Janos Sallai, Will Hedgecock, Peter Volgyesi, Andras Nadas, Gyorgy Balogh, and Akos Ledeczi. 2011. Weapon classification and shooter localization using distributed multichannel acoustic sensors. *Journal of Systems Architecture* 57, 10 (2011), 869–885.

[64] Ralph Schmidt. 1986. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation* 34, 3 (1986), 276–280.

[65] Prem Seetharaman and Stephen P Tarzia. 2012. The hand clap as an impulse source for measuring room acoustics. In *Audio Engineering Society Convention 132*. Audio Engineering Society.

[66] Souvik Sen, Romit Roy Choudhury, Bozidar Radunovic, and Tom Minka. 2011. Precise indoor localization using PHY layer information. In *Proceedings of the 10th ACM Workshop on Hot Topics in Networks*. ACM, 18.

[67] Shahram Shahbazpanahi, Shahrokh Valaee, and Mohammad Hasan Bastani. 2001. Distributed source localization using ESPRIT algorithm. *IEEE Transactions on Signal Processing* 49, 10 (2001), 2169–2178.

[68] Tie-Jun Shan, Mati Wax, and Thomas Kailath. 1985. On spatial smoothing for direction-of-arrival estimation of coherent signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 33, 4 (1985), 806–811.

[69] Sheng Shen. 2019. *Actively exploiting propagation delay for acoustic systems*. Ph.D. Dissertation. University of Illinois at Urbana-Champaign.

[70] Sheng Shen, Nirupam Roy, Junfeng Guan, Haitham Hassanieh, and Romit Roy Choudhury. 2018. MUTE: Bringing IoT to noise cancellation. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*. ACM, 282–296.

[71] Sheng Shen, Nirupam Roy, Junfeng Guan, Haitham Hassanieh, and Romit Roy Choudhury. 2018. Networked Acoustics Around Human Ears. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, 747–749.

[72] Yoke Leen Sit, Christian Sturm, Johannes Baier, and Thomas Zwick. 2012. Direction of arrival estimation using the MUSIC algorithm for a MIMO OFDM radar. In *2012 IEEE Radar Conference*. IEEE, 0226–0229.

[73] Piergiorgio Svaizer, Alessio Brutti, and Maurizio Omologo. 2012. Environment aware estimation of the orientation of acoustic sources using a line array. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 1024–1028.

[74] Honghao Tang. 2014. *DOA estimation based on MUSIC algorithm*. Technical Report. Linnaeus University, Department of Physics and Electrical Engineering. 56 pages.

[75] Gabriel Taubman and Brian Strope. 2014. Speech recognition models based on location indicia. US Patent 8,831,957.

[76] Bert Van Den Broeck, Alexander Bertrand, Peter Karsmakers, Bart Vanrumste, Marc Moonen, et al. 2012. Time-domain generalized cross correlation phase transform sound source localization for small microphone arrays. In *2012 5th European DSP Education and Research Conference (EDERC)*. IEEE, 76–80.

[77] Krishnaraj M Varma. 2002. *Time delay estimate based direction of arrival estimation for speech in reverberant environments*. Ph.D. Dissertation. Virginia Tech.

[78] Jie Xiong and Kyle Jamieson. 2013. Arraytrack: A fine-grained indoor location system. In *Presented as part of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13)*. 71–84.

[79] Faheem Zafari, Athanasios Gkelias, and Kin K Leung. 2019. A survey of indoor localization systems and technologies. *IEEE Communications Surveys and Tutorials* (2019).

[80] Xing Zhang, Lin Zhong, and Ashutosh Sabharwal. 2018. Directional training for FDD massive MIMO. *IEEE Transactions on Wireless Communications* 17, 8 (2018), 5183–5197.