AIM: Acoustic Imaging on a Mobile

Wenguang Mao, Mei Wang, and Lili Qiu
The University of Texas at Austin
{wmao,meiwang,lili}@cs.utexas.edu

ABSTRACT

The popularity of smartphones has grown at an unprecedented rate, which makes smartphone based imaging especially appealing. In this paper, we develop a novel acoustic imaging system using only an off-the-shelf smartphone. It is an attractive alternative to camera based imaging under darkness and obstruction. Our system is based on Synthetic Aperture Radar (SAR). To image an object, a user moves a phone along a predefined trajectory to mimic a virtual sensor array. SAR based imaging poses several new challenges in our context, including strong self and background interference, deviation from the desired trajectory due to hand jitters, and severe speaker/microphone distortion. We address these challenges by developing a 2-stage interference cancellation scheme, a new algorithm to compensate trajectory errors, and an effective method to minimize the impact of signal distortion. We implement a proof-of-concept system on Samsung S7. Our results demonstrate the feasibility and effectiveness of acoustic imaging on a mobile.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; **Visualization techniques**;

KEYWORDS

Acoustic imaging, SAR, autofocus, interference cancellation

ACM Reference Format:
Wenguang Mao, Mei Wang, and Lili Qiu. 2018. AIM: Acoustic Imaging on a Mobile. In *MobiSys '18: The 16th Annual International Conference on Mobile Systems, Applications, and Services, June 10–15, 2018, Munich, Germany*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3210240.3210325

1 INTRODUCTION

Motivation: The ability to image an object has profound applications to the society, such as health care, entertainment, news, and much more. With the unprecedented growth of smartphone popularity, smartphone based imaging is very appealing. While smartphone cameras are getting increasingly powerful, they still lack in many scenarios, such as imaging in darkness or under obstruction. RF based imaging is an interesting alternative. For instance, RF imaging radars [43] have been widely used to monitor weather and identify military targets. However, these radars are big, power hungry, and mechanically complex. Recently, [1, 31, 60, 65] develop pioneering RF imaging systems based on light-weight devices, such as drones, RFID arrays, and millimeter-wave transceivers. While these devices are more accessible and affordable than radars, they still require special hardware and significant effort to set up.

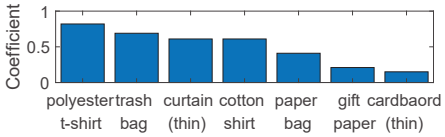


Figure 1: Penetration coefficients of sounds.

Acoustic imaging is another alternative. It complements camera based imaging in the following scenarios. First, its performance does not depend on lighting condition and it can be used in darkness. An interesting application is using acoustic imaging for indoor mapping, which is preferred to perform at night to minimize the impact of human activities. A user can also use acoustic imaging to detect obstacles on the road at night or in caves.

Second, acoustic signals can penetrate many materials as shown in Figure 1, and support imaging objects covered by these materials. With this capability, a policeman can use acoustic imaging to detect weapons under clothes, which may potentially help prevent recent shooting tragedies [56, 63].

Third, acoustic signals can propagate around obstructions through diffraction on their edges [49] or reflection from nearby furniture or walls. This capability supports imaging objects behind obstructions even if the signals cannot penetrate them directly. With this capability, a robot can use acoustic imaging to see around corner [6] and plan its movement correspondingly.

Compared with RF based approaches, acoustic imaging has two advantages. First, we can easily customize transmission signals and process received signals in software without special hardware. Thus, acoustic imaging can be implemented as a mobile app. Second, acoustic signals are much slower than RF signals. This helps achieve high image resolution, which is determined by the ratio between the signal propagation speed and bandwidth [28]. To achieve the same resolution of acoustic imaging using 10 KHz bandwidth, an RF based system needs around 9 GHz bandwidth!

Challenges: While acoustic imaging is attractive, enabling it on a mobile involves significant challenges. Ultrasound medical imaging uses transducer arrays to send and receive signals to generate high-quality images. However, a smartphone has a small number of microphones and speakers, which are insufficient to form a sizable array. To solve this problem, we apply synthetic aperture radar (SAR) for imaging [12, 44]. As shown in Figure 2, we move a phone in front of the target to mimic a virtual microphone array. To realize SAR imaging on a smartphone, we should address the following challenges that are unique to smartphone based acoustic imaging:
1) Self and background interference: In addition to reflection from

MobiSys '18, June 10–15, 2018, Munich, Germany

ACM ISBN 978-1-4503-5720-3/18/06…$15.00
https://doi.org/10.1145/3210240.3210325

Figure 2: Smartphone based acoustic imaging.

the target, signals received by the microphone contain direct transmission from the speaker and background reflection. These signals overlap in time with target reflection and cause significant interference, making the image too noisy to see. In comparison, RF based imaging systems use directional antennas to limit interference.

2) Deviation from the trajectory: SAR requires the user to move a phone along a predefined path (*e.g.*, a straight line), but it is hard for a hand to exactly follow the desired trajectory. Trajectory errors translate to phase errors of received signals and significantly blur the image. Several autofocus algorithms are proposed for RF based systems to estimate and compensate for these errors [12]. Among them, Phase Gradient Algorithm (PGA) is the most effective [17, 59]. However, our experiments show that PGA cannot be directly applied to our context. A close examination reveals that PGA assumes narrow beam signals (*i.e.*, the carrier frequency is much larger than bandwidth), well separated dominant reflectors, and no quantization errors. These assumptions do not hold for mobile acoustic imaging systems due to low carrier frequency, short imaging distance, and digital signal processing.

3) Speaker and microphone distortion: To get high-quality images, we use signals with large bandwidth (10 KHz to 22 KHz). However, the frequency response of speakers and microphones on mobiles is not flat across the selected band. This is not surprising since frequencies above 15 KHz are hardly audible and not optimized. The uneven response introduces significant distortion to signals and blurs the generated images. Commodity speakers and microphones introduce such strong distortion. RF and ultrasound transceivers designed for imaging purpose do not have this problem.

Our approach: We develop an **A**coustic **I**maging system on a **M**obile, called AIM. It is based on SAR, where a user holds a mobile and swipes over a line in front of an object, as shown in Figure 2.

We address (1) by developing a 2-stage interference cancellation scheme. In the first stage, we cancel the self interference by subtracting pre-recorded direct path signals from received samples. We account for automatic gain control (AGC) of the microphone to enhance the cancellation. In the second stage, we remove the background interference by exploiting the fact that it has different propagation delay from the target reflection.

We address (2) by developing a new phase error correction algorithm called *MPGA*. MPGA consists of two major components: (i) estimating and compensating for quantization errors, which are ignored in existing SAR imaging, (ii) using stochastic methods to study received signals and estimate phase errors so that we can remove the assumption of narrow beam signals and capture the impact of closely spaced dominant reflectors.

We address (3) based on a key observation that speaker and microphone distortion blurs the image in a deterministic way. We measure their frequency response and find out the distortion pattern

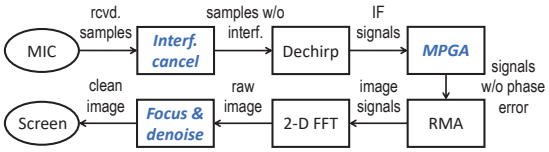


Figure 3: AIM system.

introduced by them. A distorted image is expressed by the product of that pattern and the corresponding undistorted image. We solve the undistorted image using Lasso regression. Meanwhile, this procedure also removes the noise from the image.

The processing flow of AIM is summarized in Figure 3. Upon receiving acoustic samples, AIM applies our 2-stage cancellation approach to remove interference. It then performs the standard dechirp process and uses our MPGA to compensate phase errors. Following that, it uses the standard Range Mitigation Algorithm (RMA) [12] to get an image. Finally, it applies focus and denoise procedure to remove distortion and noise from the image. To evaluate AIM, we implement a proof-of-concept system based on commercial smartphones. The experiment results show that AIM can effectively capture images for various objects with a similarity of 0.7 to 0.9 in line-of-sight (LoS), under-clothes, and in-bag scenarios. Processing delay of AIM is only 1.2 s on Samsung S7.

The contribution of this paper is summarized as follows:

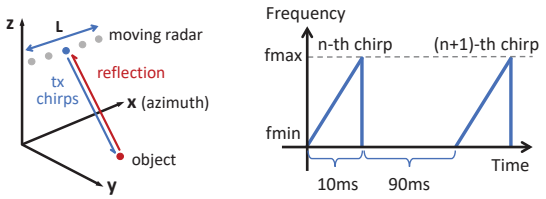
- We propose a new phase error correction algorithm MPGA for acoustic imaging under mobile contexts. It can be applied to other scenarios to remove the impact of imperfect motion.
- We develop an approach to remove distortion and noise from images. The idea is beneficial to other applications where speaker and microphone distortion is a concern, such as acoustic tracking.
- We implement an acoustic imaging system on a mobile, and demonstrate its feasibility using experiments.

2 BACKGROUND ON SAR IMAGING

Radio-frequency imaging is widely used for remote sensing applications, such as earth observation and military surveillance. The commonly used signal frequency is from 1 GHz to 40 GHz [43]. The key technique behind such imaging system is Synthetic Aperture Radar (SAR). Its main idea is moving a radar with small aperture over a long distance to emulate a large-aperture radar that helps generate images with much higher resolution. A linear SAR system is shown in Figure 4(a), where the radar moves along the x-axis (called *azimuth direction*). The total distance moved is called *synthetic aperture*, denoted as L.

During the movement, the radar sends chirps periodically, whose frequency linearly sweeps from the minimum to the maximum over time, as shown in Figure 4(b). Meanwhile, it collects the reflected signals to generate images. The separation between two chirps is large enough to ensure that all reflected signals of the current chirp are received before the next chirp is transmitted.

We use n to denote the index of transmitted chirps, called *azimuth index*, since the chirps are sent as the radar moves along the azimuth direction. For each transmitted chirp, we use k to denote the index of the received samples, called *range index*, since early samples are reflected by the objects with a shorter range to the radar. By multiplying with the transmitted chirps (*i.e.*, *dechirp*), the received samples are down-converted and stored as a 2-D data matrix $s(n, k)$,



(a) System view (b) Chirp signals

Figure 4: Synthetic aperture radar.

called *intermediate-frequency (IF) signals*. Following that, IF signals go through a series of signal processing known as *data formatting*. The obtained signals are called *image signals* and denoted by $i(n, k)$. The purpose of data formatting is to make the image signals for a point reflector located at $(x, y, 0)$ have the following expression

$$i(n,k) = Ae^{jc(xn+yk)}, \tag{1}$$

where A is the magnitude of received signals and c is a constant. Note that Equation 1 is a simplified formula based on the assumptions that the distance between the radar and target is long and the signal propagation speed is high. These assumptions are valid for RF radar applications. Refer to [12] for the complete expression of Equation 1 and its derivation. Importantly, we notice that $i(n, k)$ is a 2-D sinusoid signal. By applying 2-D Fast Fourier Transform (FFT), we can observe a spike located at (cx, cy) in the 2-D frequency space. Thus, the above procedure maps a point in the physical space (*i.e.*, the xy plane in Figure 4(a)) to a point in the 2-D frequency space. Since the mapping is linear and one-to-one, the shape of an object in the physical space is preserved in the frequency space. Hence, 2-D FFT of the image signals produces an image of the object.

The details for SAR processing, including dechirp and data formatting, are explained in [12]. In our implementation, we use Range Mitigation Algorithm (RMA) for data formatting due to its effectiveness in near field imaging [12], but our approaches are compatible with other algorithms. RMA consists of the following steps: (i) perform FFT on each column of IF signal matrix; (ii) apply the matched filter and variable substitution to convert the signals to the desired format as Equation 1; (iii) use Stolt interpolation [12] to get uniformly sampled image signals; and (iv) apply 2-D FFT on image signals to obtain the object image.

New challenges emerge when applying SAR to acoustic imaging on a mobile. First, since smartphone speakers and microphones are omni-directional, signals propagating directly from the speaker to microphone interfere with target reflection. Second, moving a mobile by hand incurs large motion errors. Existing error correction algorithms do not work because the underlying assumptions do not hold in our context. Third, smartphone speakers and microphones severely distort acoustic signals, which blurs the generated images. To address these challenges, AIM adds three new components to the SAR processing pipeline, as highlighted in Figure 3. We will explain these components in the following sections.

3 INTERFERENCE CANCELLATION

Two types of interference: Since the smartphone speaker and microphone are omni-directional, signals received by the microphone include not only desired reflection from the target, but also two types of interference: (i) direct transmission and (ii) background reflection, as shown by Figure 5. These signals overlap in time because the difference between their propagation delay is smaller than the chirp duration (*e.g.*, 10 ms in our system). Even worse, the direct transmission is 3 - 4 orders of magnitude larger than the target reflection. To minimize the impact of interference, we develop a two-stage interference cancellation scheme. In the first stage, we cancel the self interference by subtracting pre-recorded direct path signals from received samples. In the second stage, we remove the residual self interference and background interference by leveraging the fact that they have different propagation delay from the target reflection.

Stage 1: This stage aims to remove interference of the direct path between the speaker and microphone. To this end, we record the direct transmission by putting the mobile in a clean space, where no major reflectors are within one meter distance in front of the speaker and microphone. When we use the mobile to image a target object, we subtract the pre-recorded direct path signals from the received samples. We take into account synchronization and sampling offset between the pre-recorded signals and currently received samples as in [45].

In practice, simple subtraction cannot achieve optimal cancellation. Our key observation is that the automatic gain control (AGC) in the microphone [16] normalizes received signals such that the highest magnitude is close to 1. The AGC gains are slightly different under various environments, since received signals contain different background reflection. This makes the direct transmission scaled differently from recordings to recordings. Since it is orders of magnitude larger than the target reflection, a small scaling difference will lead to significant residual interference. To address this issue, we find a scaling coefficient c for each chirp period that minimizes $||\mathbf{S} - c\mathbf{S}_d||$, where $\mathbf{S}$ represents received samples in the current period and $\mathbf{S}_d$ denotes pre-recorded direct signals. The intuition behind the optimization is that when c exactly compensates the scale difference between S and S_d, the direct path signals will be fully removed and the magnitude of remaining signals is minimized. Figure 6 plots the optimal c over time in a real trace. Once c is determined, we subtract $c\mathbf{S}_d$ from $\mathbf{S}$ to remove the direct transmission. Our evaluation shows that we can cancel 30 dB interference without scaling the pre-recorded signals and 36 dB with scaling.

Stage 2: The previous stage removes most of the direct path interference. However, there are still some residuals due to imperfect synchronization and scaling between received samples and pre-recorded signals. This stage aims to remove residual direct transmission and background reflection. For this purpose, we exploit different propagation delay of these signals from that of target reflection, as shown in Figure 5. The signals after the first-stage cancellation can be described using the well-known multipath channel model [57]:

$$y[n] = \sum_{i \in U_1} h_i x[n-i] + \sum_{j \in U_2} h_j x[n-j],$$

where x denotes the transmitted signals, and h stands for the channel taps. U_1 includes the indices such that $d_1 < i \cdot t_s \cdot v_s < d_2$, where t_s is the sampling interval, v_s is the sound speed, and $[d_1, d_2]$ is the range that the target object falls into. Thus, the first summation in the above equation contains the target reflection.

Figure 5: Received signals.

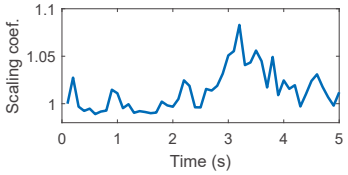


Figure 6: Scaling coefficients.

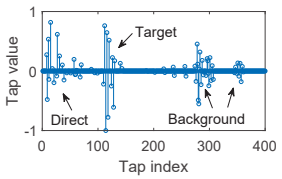


Figure 7: Estimated channel taps.

U_2 includes all other indices, and the second summation consists of residual direct transmission and background reflection. Since both y and x are known, we apply Least Square channel estimation [50] to determine channel taps, *i.e.*, finding $\{h_i\}$ such that $\sum_n (y[n] - \sum_{i\in(U1\cup U2)} h_i x[n-i])^2$ is minimized. Figure 7 shows an example of estimated channel taps from real traces. Once the taps are determined, we remove the interference by subtracting $\sum_{j\in U_2} h_j x[n-j]$ from $y[n]$. As shown in Section 8, this stage can cancel additional 5 dB interference.

Our two-stage strategy is the key to achieve large cancellation. Stage 1 is needed to remove the direct path signals. Otherwise, strong direct transmission will overwhelm the target and background reflection, which severely degrades the channel estimation on the corresponding taps and cancellation performance. Stage 2 is necessary to remove the remaining background interference.

4 MPGA

In this section, we first introduce PGA and its limitations in our context, and then present our MPGA to address these issues.

4.1 PGA

The phase error caused by imperfect motion is one of the most challenging issues in SAR. When the motion deviates from the desired path (*e.g.*, a straight line) by δ_d, the received signals experience a phase shift equal to $2 \cdot 2\pi \frac{\delta_d}{\lambda}$, where λ is the signal wavelength and a multiplier of 2 accounts for round-trip propagation. When the radar moves over the whole synthetic aperture, the motion error can be expressed as $\delta_d(n)$, where n is the azimuth index. Let $e(n)$ denote the phase error introduced by $\delta_d(n)$.

To remove phase error $e(n)$, several phase correction algorithms are developed [7, 12, 33]. Among them, Phase Gradient Algorithm (PGA) is considered the most effective [17, 59]. PGA is designed to remove the second or higher order phase errors with respect to n since a constant phase error has no impact on the image and a linear phase error only results in a shift to the image.

The basic idea for PGA is explained as follows. Equation 1 models the image signals without motion errors. As discussed above, when the errors are present, received signals experience extra phase shifts (*i.e.*, $e(n)$). These phase shifts remain in the signals after dechirp and data formatting. As a result, the image signals $i(n,k)$ under imperfect motion are given by $Ae^{[cxn+cyk+e(n)]}$ [12]. For simplicity, we let $i_k(n)$ denote the k-th column of the image signals, ω denote cx, and ϕ denote cky. Thus, we have

$$i_k(n) = i(n,k) = Ae^{j(\omega n+\phi+e(n))}. \tag{2}$$

PGA estimates the phase error $e(n)$ based on i_k. Specifically, we first apply FFT to derive the spectrum of i_k, and then identify the frequency component with the maximum magnitude, denoted as $\tilde{\omega}$. Note that $\tilde{\omega} \approx \omega$. After circularly shifting $i_k(n)$ by $\tilde{\omega}$ in the frequency domain, the time domain signal becomes $l_k(n)$ and is approximated by $Ae^{j(\phi+e(n))}$. The derivative of $l_k(n)$ is given by $\dot{l_k}(n) = jA\dot{e}(n)e^{j(\phi+e(n))}$. Then the derivative of phase error $e(n)$ can be obtained by $\dot{e} = \frac{\mathrm{Im}(\dot{l_k} l_k^*)}{l_k l_k^*}$, where $\mathrm{Im}(x)$ is the imaginary part of x, and l_k^* denotes the conjugate of l_k. The previous equation holds for any k. Therefore, we can use all available k for estimation to improve the accuracy as [17]:

$$\dot{e} = \frac{\sum_k \mathrm{Im}(\dot{l_k} l_k^*)}{\sum_k l_k l_k^*}. \tag{3}$$

The phase error $e(n)$ can be estimated by integrating its derivative. The detailed derivations for PGA can be found in [12, 17, 59].

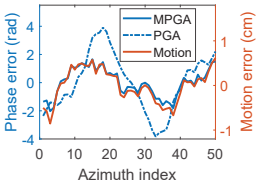
4.2 Limitations of PGA

The effectiveness of PGA depends on whether Equation 2 holds. Equation 2 requires four assumptions, which do not hold in smartphone based acoustic imaging systems.

Supporting only narrow beam signals: First, it assumes narrow beam signals: the carrier frequency is much larger than the bandwidth [55]. This assumption easily holds for RF based radars and ultrasound based sonars. The former uses GHz carrier frequency and tens of MHz bandwidth [12, 43], and the latter uses MHz carrier frequency and tens of KHz bandwidth [24]. Without this assumption, the motion error interacts with ω and ϕ in Equation 2 and we cannot separate it as an individual phase term in image signals.

However, the assumption does not hold in our context, since the highest frequency supported by smartphone speakers is around 20 KHz. To provide high image resolution, our system uses frequencies from 10 KHz to 22 KHz. In this case, the carrier frequency is 16 KHz and the bandwidth is 12 KHz, which clearly does not satisfy the narrow beam requirement. Thus, directly applying PGA cannot effectively remove phase errors in our system.

To illustrate that, we simulate imaging a point with acoustic signals. We use simulation in this section because we need to control the presence of phase errors to demonstrate their impact on imaging. For simulation, we generate synthetic received signals for a point reflector, and apply the imaging algorithms to produce images. Also, we inject random motion errors as the phone moves across the synthetic aperture. These motion errors translate to phase errors linearly. We observe that the estimated phase errors by PGA (dashed lines) do not match with the ground-truth motion errors (solid lines), as shown in Figure 8(a). Figure 8(b) shows the image without phase error correction, while Figure 8(c) shows the image using PGA correction. Figure 10(b) shows the ground truth for imaging a point. As we can see, due to inaccurate phase error estimation, the image with PGA correction is severely blurred. These results indicate that PGA is insufficient for our purpose.



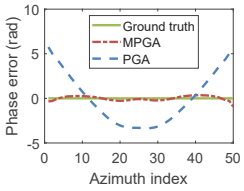
(a) Error estimation

(b) No correction

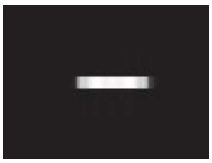
(c) PGA correction

(d) MPGA correction

Figure 8: (a) Phase errors due to imperfect motion; (b)(c)(d) images with different phase error correction.

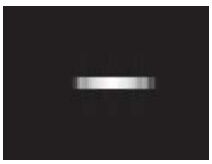


(a) Error estimation



(b) Ground truth image

(c) PGA correction



(d) MPGA correction

Figure 9: (a) Phase error estimation for array reflectors; (b)(c)(d) images with different phase error correction.

Supporting only isolated reflectors: Second, the derivation of Equation 3 assumes that the target object is a point. This is the case when there is only one dominant reflector in the imaged region, whose reflection is much stronger than other surrounding objects [17]. When there are multiple dominant reflectors, as long as there is a reflector well separated from the others, we can first use filtering in frequency domain to isolate it and then apply PGA. This works well for remote sensing applications [12, 43].

In comparison, we focus on imaging an object from a short distance (*e.g.*, 0.5 m). In this case, the object is treated as an array of *closely spaced homogeneous reflectors* (or *array reflectors*). Filtering does not work in this case. Instead, we need to generalize PGA to model phase errors in received signals for array reflectors.

Ignoring secondary phase terms: Third, PGA assumes negligible impact of secondary phase terms (*i.e.*, non-linear phase terms with respect to (x, y) omitted in Equation 1 [59]). These terms are inversely proportional to the signal speed and imaging distance [12]. Due to slow acoustic propagation and short imaging distance, these secondary terms cannot be ignored in our case. To correctly compensate phase errors, we need to minimize the impact of these secondary phase terms. It is especially challenging to determine their impact based on the received signals of array reflectors.

To illustrate that, we simulate imaging a horizontal bar with the length equal to 20 cm. The mobile scans over synthetic aperture without any motion error. As a result, the ground truth phase error should be zero for all azimuth indices. As shown in Figure 9(a), PGA significantly over-estimates the phase error since the impact of secondary phase terms is not considered. Therefore, the image with PGA phase correction in Figure 9(c) shows significant distortion, compared with the ground truth shown in Figure 9(b).

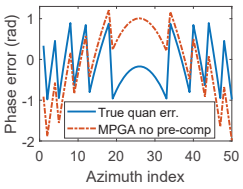
Quantization errors: In addition to the motion error, which is the major source for the phase error in received signals, our system also incurs the quantization error for the following reason. For SAR processing, the received samples are first shifted forward by $2R_a/v_s$, where v_s is the signal speed and R_a is the distance between the center of imaged region and the current position of the radar assuming no motion error. The shifted samples are then multiplied with the transmitted chirp and pass through a low-pass filter. This process is called *dechirp* [12] and required by any chirp based SAR system. For radars, shifting samples is performed in the analog domain and is precise. However, in our case, we have to shift received samples in the digital domain, since we use a built-in smartphone microphone. Thus, we can only shift the samples by multiple sample intervals. Given the sample interval t_s, the shift is $t_s \cdot \text{round}(2R_a/(v_s t_s))$. The presence of quantization errors introduces additional phase errors, which is not considered in PGA.

4.3 MPGA

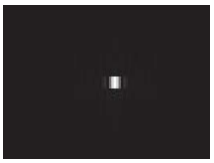
Overview: To correct phase errors of received signals for mobile acoustic imaging, we develop a new algorithm, called *MPGA*. MPGA advances PGA in the following ways. First, it estimates quantization errors due to the low sampling rate of acoustic signals. Second, it uses IF signals instead of image signals to estimate motion errors so that we can remove the assumption of narrow-beam signals. Third, it explicitly takes array reflectors and secondary phase terms into account when estimating motion errors. In this way, MPGA can effectively support acoustic imaging in a mobile context.

Removing quantization errors: MPGA first removes the phase errors introduced by quantization. We make an important observation: these errors can be determined given a specific synthetic trajectory, because they only depend on $2R_a/v_s$. Therefore, we calculate these phase errors in advance and remove them from received signals. To compute these errors, we simulate the SAR imaging with and without quantization for the given trajectory, and compare image signals in two cases to derive the errors. For a given trajectory, these errors are only calculated once. They are cached to avoid real-time computation.

Figure 10(a) shows the ground-truth phase errors introduced by quantization (solid line) versus those estimated using MPGA without removing them in advance (dotted line). The latter essentially treats quantization as part of motion errors. As we can see, the phase errors caused by quantization have sharp changes. These changes lead to singularities in the derivatives of phase errors. Since MPGA relies on the derivatives to estimate phase errors, the presence of singularities has a negative impact on the estimation accuracy. This

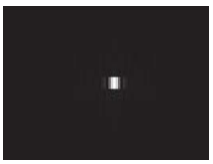


(a) Error estimation



(b) Ground truth image

(c) MPGA w/o pre-comp



(d) MPGA correction

Figure 10: (a) Phase errors due to quantization; (b)(c)(d) images with different phase error correction.

explains the difference between the ground truth and estimated values in Figure 10(a), and justifies the need to remove the quantization phase errors in advance. Figure 10(b), (c), and (d) show the images for a point reflector without quantization errors (*i.e.*, ground truth), using MPGA but without compensating quantization errors in advance, and MPGA, respectively. As we can see, the image with MPGA is much closer to the ground truth.

Estimating motion errors: Next, MPGA estimates and removes phase errors induced by imperfect motion. Different from PGA, which uses image signals for estimation [59], MPGA uses IF signals for two reasons. First, using IF signals removes narrow-beam assumption while still allows us to transform the signals to a desired format as Equation 1. Second, we can explicitly handle secondary phase terms, because they can be easily derived from IF signals.

Similar to PGA, we manipulate received samples to get an expression similar to Equation 2 to estimate the derivatives of phase errors. The challenge lies in supporting array reflectors and determining the impact of the secondary phase terms. In Section 4.4, we prove the following theorem:

Theorem 1. *The motion error $R_e(n)$ can be estimated by $E(n) - R_s(n)$, where n is the azimuth index, $R_s(n)$ is the offset introduced by the secondary phase terms and is a 2-order polynomial of n, $E(n) = \int_n \frac{\sum_k \mathrm{Im}(\frac{1}{K_R^k}\dot{s}_k s_k^*)}{\sum_k s_k s_k^*}$, s_k is the k-th column of IF signal matrix $s(n,k)$, $\dot{s}_k$ is its derivative, s_k^* is its conjugate, $\mathrm{Im}(x)$ is the imaginary part of x, and K_R^k is a constant depending on k.*

Based on Theorem 1, we first compute $E(n)$ based on IF signals. To calculate $R_s(n)$, we need to know its expression. However, it depends on the shapes of imaged objects and is not known in advance. Theorem 1 indicates that $R_s(n)$ is a 2-order polynomial of n. Thus, we only need to figure out its second order coefficient, because the first order and constant phase errors have no impact on imaging. We observe that the hand motion error $R_e(n)$ is dominated by noise-like fluctuation around zero when intentionally moving along a straight line. If fitting such a pattern using a 2-order polynomial, the second order coefficient is usually close to zero, as shown in Figure 11. As a result, we obtain the second order coefficient of

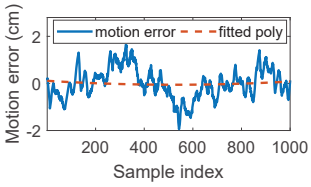


Figure 11: Motion errors in a real trace.

$R_s(n)$ by fitting $E(n)$ using a 2-order polynomial, since the second order coefficient of their difference is zero. Then, we compute $E(n) - R_s(n)$ to get the motion error $R_e(n)$. Once $R_e(n)$ is estimated, the corresponding phase errors can be determined and removed from the signals. As shown in Figure 8, 9, and 10, MPGA gives clear images of the target.

4.4 Proof of Theorem 1

The development of MPGA depends on the correctness of Theorem 1. In this section, we provide the key steps about how the theorem is derived. Refer to [39] for the detailed proof.

PGA uses image signals to estimate phase errors. Instead, we use IF signals to remove the assumption of narrow-beam signals and explicitly handle secondary phase terms. According to [12], the k-th column of IF signals is given by:

$$s_k(n) = s(n,k) = \sum_i e^{j\{\omega_i^k n + \phi_i^k + K_R^k R_e(n) + \theta_i^k(n)\}},$$

where n is the azimuth index, k is the range index, and i is the index for reflectors, ω_i^k, ϕ_i^k, and K_R^k are constants. $R_e(n)$ is the motion error. The phase errors introduced by $R_e(n)$ vary with k since K_R^k is different. Therefore, instead of directly estimating phase errors like PGA, we first determine motion errors and then scale them to get phase errors for various k. $\theta_i^k(n)$ captures the secondary phase terms, which are non-linear with respect to n and can be approximated by a 2-order polynomial of n [12].

The intuition of our proof is to derive a formula similar to Equation 3. That is, we establish the relationship between $R_e(n)$ and $[\sum_k \mathrm{Im}(\frac{1}{K_R^k}\dot{s}_k s_k^*)]/[\sum_k s_k s_k^*]$. The challenge is that s_k in our case is much more complicated than l_k in PGA derivation, since we consider array reflectors and secondary phase terms. To solve this problem, we define $\Theta_i^k(n) = \omega_i^k n + \phi_i^k + \theta_i^k(n)$. Since $\Theta_i^k(n)$ is a phase, we only care about its remainder divided by 2π. Our key observation is that $\Theta_i^k(n)$ varies significantly for different i (*e.g.*, 0 to 60 rad in our case) so that its remainder widely distributes over the range of $[0, 2\pi]$. To simplify s_k, we treat $\Theta_i^k(n)$ as a uniform random variable over $[0, 2\pi]$, and Θ_i^k and Θ_j^k are independent when $i \neq j$. Then s_k is given by

$$s_k(n) = \sum_i e^{j\{K_R^k R_e(n) + \Theta_i^k(n)\}}.$$

Based on this transform, we can approximate $\sum_k s_k s_k^*$ using its expectation if the standard deviation is small. We show that

$$\mathrm{E}[\sum_k s_k s_k^*] = KM \text{ and } \mathrm{Var}[\sum_k s_k s_k^*] = KM(M-1),$$

where K is the number of range indices and M is the number of reflectors. When K is large (*e.g.*, 592 in our implementation), $E[\sum_k s_k s_k^*]$ is much larger than the standard deviation $\sqrt{KM(M-1)}$.

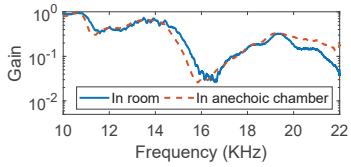


Figure 12: Combined freq. response of speaker & mic.

Thus, it is reasonable to approximate $\sum_k s_k s_k^*$ with its expectation. By applying a similar trick, we show that

$$\sum_k \text{Im}(\frac{1}{K_R^k}\dot{s}_k s_k^*) \approx \dot{R_e}(n)\sum_k s_k s_k^* + \sum_k \sum_i \frac{1}{K_R^k}\dot{\Theta_i^k}(n),$$

Thus, we have

$$\frac{\sum_k \text{Im}(\frac{1}{K_R^k}\dot{s}_k s_k^*)}{\sum_k s_k s_k^*} \approx \dot{R_e}(n) + \frac{\sum_k \sum_i \frac{1}{K_R^k}\dot{\Theta_i^k}(n)}{KM}.$$

Let $E(n)$ denote the integral of the above equation. Then,

$$E(n) = \int_n \frac{\sum_k \text{Im}(\frac{1}{K_R^k}\dot{s}_k s_k^*)}{\sum_k s_k s_k^*} = R_e(n) + \frac{\sum_k \sum_i \frac{1}{K_R^k}\Theta_i^k(n)}{KM}.$$

Let $R_s(n)$ denote the last term of the above equation. Then $E(n) = R_e(n) + R_s(n)$. Since Θ_i^k is a 2-order polynomial of n for any k and i, $R_s(n)$ is also a 2-order polynomial.

5 FOCUS AND DENOISE

Blur and noise in images: The image resolution of AIM is determined by signal bandwidth. To achieve high resolution, we use chirps sweeping from 10 KHz to 22 KHz. However, the built-in speaker/microphone on a mobile has different gains at these frequencies, as shown in Figure 12. We see that there is 16 dB difference between the minimum and maximum gains. As a result, acoustic signals are distorted. The distortion causes the envelop of a received chirp to change over time. Therefore, the signal magnitude A in Equation 1 depends on the range index k, since k indicates the time order of received samples. In this case, Equation 1 is not a standard 2-D sinusoid, and its FFT is not a delta function in the 2-D frequency space and experiences certain spread over the vertical direction (corresponding to k). Hence, when imaging a point, we see a blurred strip, instead of a clear point, as shown in Figure 13(a).

Also, we observe noise distributed over the whole image in Figure 13(a). This is due to imperfect interference cancellation and the presence of environment noise. To get high-quality images, we need to eliminate blur (*focus*) and remove noise (*denoise*).

Frequency response measurement: To minimize the blur, we need to know the gains of the built-in speaker and microphone at various frequencies, *i.e.*, their *frequency response*. Since we only care about the aggregate distortion of the speaker and microphone, we measure their combined frequency response. The procedure is outlined as follows. First, we place the mobile on a table, play the acoustic signals with the speaker, and record them with the microphone. Then, we place a 2cm×2cm cardboard at 20 cm distance in front of the speaker and microphone, and calculate the difference between the recorded signals with and without the cardboard to remove all echoes except the one from the cardboard. Finally, we compare the obtained signals with the transmission signals to derive the frequency response. This procedure requires no special equipment and can be easily repeated by users on their own phones.

To evaluate the accuracy of this approach, we perform another measurement in the anechoic chamber. We use two identical phones, which have same speakers and microphones. We let one phone play acoustic signals and the other record them. Since there is no echo in the chamber, the recorded signals only experience speaker and microphone distortion. By comparing the recorded signals with the transmission signals, we compute the frequency response. As shown in Figure 12, the anechoic chamber measurement matches the in-room measurement, which indicates the high accuracy of in-room measurement.

Focus and denoise using Lasso regression: Once the frequency response is known, a natural approach to cancel its effect is to compensate the acoustic signals with the inverse response before they are transmitted. However, since the frequency response has a deep notch at 16 KHz as shown in Figure 12, we have to significantly reduce the power of other frequency components to get flat gains. This method results in severe reduction in received signal strength and degradation on image quality.

To remove the blur, we make the following important observation: the blur introduced by speaker and microphone distortion is deterministic, since their frequency response is unchanged. Therefore, we can include the measured frequency response in our simulation, and determine how the image of a single point spreads due to the distortion. Such spread is called *point response*. There are two observations about the point response of our system. First, it is 1-D because the distortion makes the 2-D FFT of Equation 1 spread over the vertical direction as discussed earlier. Second, it is invariant to the position of the imaged point. Since any object consists of a set of points, its generated image is the superposition of shifted versions of point response. Mathematically, we have

$$I = RX,$$

where X is the object image without the blur, and each of its non-zero element represents a point on the object. R captures the effect of distortion. Each R's column is a shifted version of the point response. I is the generated image under the distortion. Since both I and R are known, we can solve X to get a clear image of the object. To this end, we use Lasso regression [25] to find X that minimizes

$$|I - RX|^2 + \lambda|X|,$$

where λ is the regularization parameter. Lasso regression not only removes the blur effect, but also helps suppress the noise. This is because the noise distributes randomly and does not match the pattern of point response, and the regularization term prevents X from overfitting the noise pattern. The effectiveness of Lasso regression depends on the selection of λ. If it is too small, X will overfit the noise in the raw image I. If it is too large, some weak reflectors will be treated as noise and removed from the image. Based on our experiments, we select λ equal to 0.01 to balance these factors. Figure 13(b) shows the image after applying our approach.

The above approach is conceptually close to the CLEAN algorithms [22, 34] in radio astronomy. They are invented to remove artifacts introduced by sidelobes of phased array antennas. In our context, the main reason for degraded image quality is the speaker

(a) w/o focus & denoise (b) focus & denoise

Figure 13: Imaging a 2cm×2cm cardboard.

and microphone distortion. Hence, our approach captures and corrects the impact of such distortion. The CLEAN algorithms use deconvolution to obtain a clean image, while we use Lasso regression so that our approach can adapt to different ambient noise levels by tuning the regularization parameter.

6 DISCUSSION

We discuss potential ways to further improve the usability of AIM.

Imaging range: Based on the experiments, the imaging range of our system is about 0.6 m. A simple way to increase the range is to increase the speaker volume. In fact, the volume of our phone (Samsung S7) is lower than the average volume of phone speakers [54]. Many devices support 10 - 17 dB higher volume [30]. Another way to improve the range is to exploit multiple speakers and microphones available on a mobile, which can form a MIMO imaging radar to improve SNR [19]. More sophisticated interference cancellation is also helpful to reduce interference, making it easier to detect weak received signals.

Image resolution: The image resolution of AIM depends on two factors. For horizontal dimension, the resolution is inversely proportional to the length of synthetic aperture L. In our system, we choose L equal to 25 cm to balance the scan effort and image quality. Increasing L is helpful to improve the resolution. For vertical dimension, the resolution is inversely proportional to the signal bandwidth. The built-in speaker on our phone can send signals up to 22 KHz, while external miniature speakers (*e.g.*, ones used in headphones) may support up to 43 KHz [5]. One can attach such miniature speaker on the mobile to replace the built-in one. This greatly helps improve the image resolution while remaining compact and easy to move.

3-D imaging: Our system can be extended to support 3-D imaging by using two microphones. The phase difference between them provides the position information on the third dimension [44].

7 IMPLEMENTATION

We implement our system on an off-the-shelf smartphone (Samsung Galaxy S7 unlocked version). We use its built-in speaker to play the pre-generated audio file to transmit acoustic signals. The signals are linear chirps whose frequency sweeps from 10 KHz to 22 KHz during 10 ms. The interval between two consecutive chirps is 90 ms to minimize interference between them.

We use the built-in microphone at the bottom of our mobile to receive signals. The sampling rate is 48 KHz, which is supported by most phones. The separation between the speaker and microphone is only 5 mm, and hence the interference caused by direct transmission is significant. We develop an Android app to process received signals and generate images. We use NDK to implement our processing algorithms to maximize the efficiency. We use FFTW package [20] for FFT operations. We use GNU Scientific Library (GSL) for other mathematical operations in our algorithm.

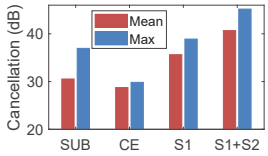
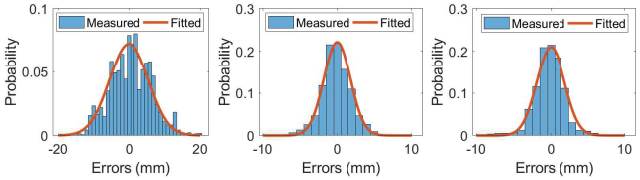


Figure 14: Interference cancellation performance.



(a) x-axis (b) y-axis (c) z-axis

Figure 15: Non-linear hand motion errors.

8 EVALUATION

In this section, we evaluate each component of AIM and its overall performance. For experiments, we put the target object on a stand, and remove objects within 1 m from the target. A user stands in front of the target, holds a phone in hand, and swipes 25 cm along a straight line, as shown in Figure 2. We put two markers separated by 25 cm on the user's clothes to provide rough reference positions for the start and end of synthetic aperture. The swipe takes 5 s. A timer is displayed on the UI of our app to provide time reference. The distance between the phone and target is 0.4 m. Although some rough reference is provided, the user hand movement cannot exactly follow the desired trajectory and speed. Deviation in either of them causes motion errors. We use MPGA to correct these errors.

8.1 Micro Benchmark

Interference cancellation: To measure the performance of interference cancellation, we collect five traces without placing an object in our experiment setup. We quantify the interference cancellation by comparing the signal strength before and after the cancellation. We compare our approach with two common digital interference cancellation methods: 1) subtract the pre-recorded direct transmission without considering AGC; 2) directly estimate channel taps using Least Square and remove ones outside the target region (0.4 m - 0.6 m from the mobile).

As shown in Figure 14, the subtraction based method (*SUB*) cancels 30 dB interference. The channel estimation based method (*CE*) only cancels 29 dB because the presence of direct transmission overwhelms the target and background reflection and degrades the estimation accuracy on the corresponding channel taps. Our first stage cancellation (*S1*) is based on subtraction based method but takes AGC into account. It achieves 6 dB higher cancellation than subtraction based method. Our 2-stage cancellation scheme (*S1+S2*) cancels 41 dB in total by removing residual interference using the channel estimation.

Hand motion errors: To quantify hand motion errors, we ask 5 users to swipe a mobile by hand as described in our experiment

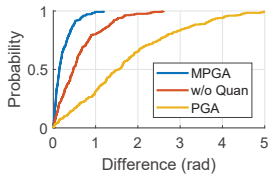
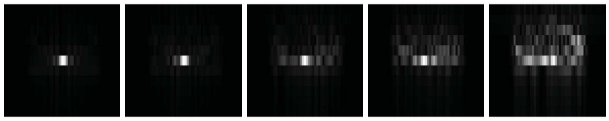


Figure 16: Difference of actual and estimated phase err.



(a) 0 rad (b) 0.2 rad (c) 0.4 rad (d) 0.8 rad (e) 1.6 rad

Figure 17: Images with various amounts of phase errors.

setup. Each user has 5 minutes to get familiar with our system and then swipes a mobile for 5 times. We use two cameras (Microsoft Q2F-00013 [15]) to record hand motions by tracking two green markers on each user's hand. One camera tracks motions in the x and z axes, and the other tracks those in the y-axis. The tracking is implemented based on color filtering using OpenCV 3.0 [10]. The distance between the camera and hand is 40 cm. We create favorable lighting condition so that camera based tracking is accurate. The motion errors are derived by comparing the tracked coordinates with the desired trajectory. Also, we remove the linear and constant motion errors by fitting the coordinates over time using a linear model. As discussed, these errors have no impact on image quality. The non-linear motion errors for each axis are shown in Figure 15. The distribution of these errors can be modeled by Gaussian functions with zero mean, and the standard deviations for three axes are 6 mm, 2 mm, and 2 mm. We observe that the x-axis (azimuth direction) has the largest error since users cannot keep a desired speed when swiping along this direction and speed errors translate to motion errors. The non-linear motion errors are within 1 cm in most cases. This is as expected because it is not difficult for users to keep a linear hand motion over a short distance (*i.e.*, 25 cm).

MPGA: Next, we evaluate MPGA by comparing the actual and estimated phase errors using MPGA. Although camera based tracking is acceptable for capturing motion errors, its accuracy is insufficient to derive phase errors, since a small tracking error (*e.g.*, 1 mm) translates to a large phase change (*e.g.*, 0.6 rad). Thus, we generate synthetic motion traces and received signals for imaging rectangle, triangle, and diamond shapes. The motion errors follow the distributions as shown in Figure 15. In this way, we have the ground truth for the motion trajectory and corresponding phase errors. We apply MPGA on the synthetic traces to estimate the phase errors. For comparison, MPGA without compensating quantization errors (w/o Quan) and PGA are also evaluated, as shown in Figure 16. We observe that MPGA yields the lowest median estimation error: 0.2 rad. Without compensating quantization, the error increases to 0.5 rad. PGA has the median error of 1.6 rad, which indicates it is ineffective in our context. To illustrate the impact of the remaining phase errors not corrected by these algorithms, we show the images of a point with different amounts of phase errors in Figure 17. We see that the image quality is degraded when the average phase error is larger than 0.4 rad.

Ongoing proc.		Post processing			
Dechirp	Intf. cancel	MPGA	RMA	FFT	Focus
0.005 s	0.04 s	0.16 s	0.46 s	0.35 s	0.26 s

Table 1: Running time of AIM.

The effect of MPGA on imaging a real object is shown in Figure 18, where the target is a bar-shape cardboard as shown in Figure 18(a). Without applying MPGA, the images are severely distorted as shown in Figure 18(b) and (c). Figure 18(e) shows that the image quality is significantly improved when MPGA is applied.

Focus and denoise: Figure 18 also shows the effectiveness of our focus and denoise algorithm. As discussed in Section 5, speaker and microphone distortion blurs the image along the vertical direction. The effect can be easily observed when imaging a horizontal bar, as shown in Figure 18(d). By applying our focus and denoise algorithm, the blur effect is removed and the noise in the image is effectively suppressed, as shown in Figure 18(e).

We repeat the above experiment using various phones (Samsung S7 AT&T version and S8), which have different speaker and microphone frequency response from our main phone (S7 unlocked version), as shown by Figure 12 and 19. Although various response causes different distortion on images, our algorithm can remove the blur and provide clean images on both phones as shown by Figure 19(b) and (c). This demonstrates the robustness of our focus and denoise algorithm and shows that our system is general and applicable to different phones.

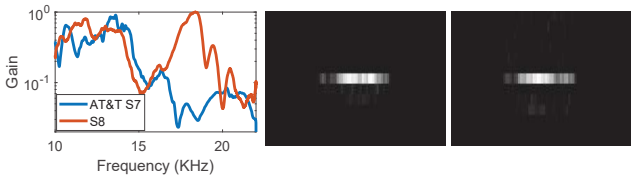
Resolution: We evaluate the resolution of our system by imaging a point-like object and measure its spread. Since the wavelength of our signals is around 2 cm, we use a 2 cm×2 cm cardboard as the target. A smaller object can cause signals to traverse it via diffraction. The generated image is shown in Figure 13(b). The spread, which is computed from the pixel with the maximum magnitude to the pixel with 3 dB degradation, is 1.8 cm in the vertical direction and 1.9 cm in the horizontal direction. This resolution is sufficient to capture the shapes of many daily objects.

Audibility: To achieve high image resolution, we use signals from 10 KHz to 22 KHz. To quantify the audibility, we measure the loudness of our signals using a software sound meter [4]. At 0.5 m distance from the speaker and using the same volume as other experiments (50% of the maximum), the loudness is around 49 dB. For reference, the loudness of ambient sounds in our lab is 35 dB. while that of human talking is 62 dB. Also, our signals are only played when the mobile scans over synthetic aperture, which takes about 5 s to complete. Thus, the impact of the sound coming from our imaging system is small.

Running time: Table 1 shows the running time of major components in AIM. Dechirp and interference cancellation are performed every time when reflected signals of a transmitted chirp are received. Their processing time for one chirp period is 5 ms and 40 ms, respectively. Their total time is within the interval between two consecutive chirps (90 ms). As a result, once we finish moving the mobile over the synthetic aperture, dechirp and interference cancellation are also done and have no impact on processing delay. MPGA, RMA formatting, FFT, and focus and denoise are applied after all signals are collected. In total, they take 1.2 s to complete. As a result, the processing delay of AIM is 1.2 s.

(a) Object (b) No focus or MPGA (c) Only focus (d) Only MPGA (e) Focus and MPGA

Figure 18: Imaging a horizontal bar: (a) ground truth; (b)(c)(d)(e) images with applying MPGA / focus or not.



(a) Response (b) AT&T S7 (c) S8

Figure 19: Imaging a horizontal bar with different phones.

8.2 System Benchmark

We evaluate the performance of AIM using various objects. To measure the image quality, we consider the following metrics:

- Szymkiewicz-Simpson similarity $s = S_c/\sqrt{S_i S_g}$, where S_i and S_g are the object areas in the generated image and ground truth image, respectively, and S_c is their intersection. The area is calculated by checking if the magnitude of each pixel is greater than a threshold 0.3 (the maximum pixel is normalized to 1). The pixel with smaller magnitude indicates no reflection from the corresponding position and is not counted as a part of the object. The similarity value equal to 1 indicates perfect matching.
- Height error h_e, defined as $|h_i - h_g|/h_g$, where h_i is the object height in our image and h_g is the ground truth.
- Width error w_e, defined similarly to the height error.
- Ratio error r_e, defined as $|r_i - r_g|/r_g$, where r_i and r_g are estimated and actual ratios between the height and width.

To smooth the object boundary, the images are linearly interpolated from 60×24 pixels to 160×120 pixels. To remove noise out of the interested region, we apply a binary mask on the images. The size of the mask is 40 cm×25 cm. The image is displayed using gray scale, where brighter points indicate higher magnitude. Each experiment is repeated for 5 times. We report the average performance of each experiment and show the images with the median Szymkiewicz-Simpson similarity.

Imaging shapes: We image the cardboards with various shapes under Line-of-Sight (LoS) scenarios, including rectangle (25 cm×18 cm), diamond (23 cm×17 cm), triangle (24 cm×17 cm), circle (19 cm×19 cm), and hollow rectangle (28 cm×18 cm). Their pictures (ground truth) are shown in the first row of Figure 20, and the images generated by AIM are shown in the second row. We observe the object shapes in the generated images match the ground truth well, and the similarity metrics are from 0.72 to 0.89 for different shapes, as shown in Table 2. The height, width, and ratio errors are also small, ranging from 0.01 to 0.11. These results indicate the effectiveness of AIM.

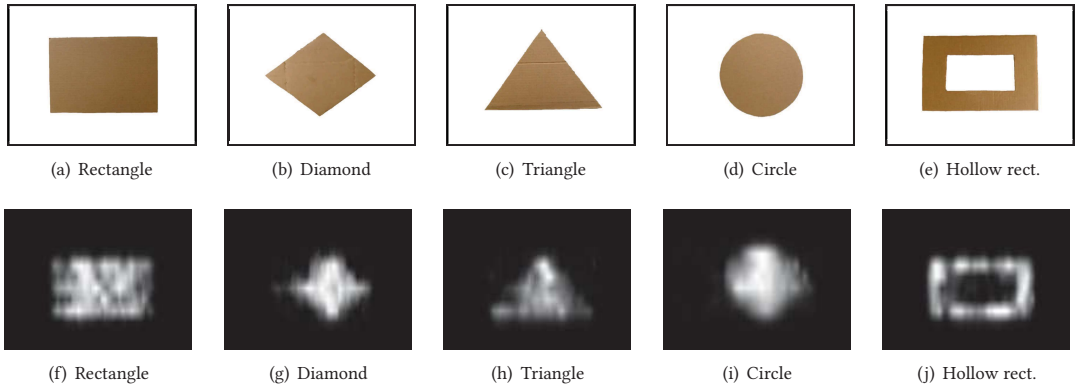
Imaging weapons: Weapon detection is a potential application of our techniques. In this experiment, we evaluate the performance of imaging a toy gun (18 cm×14 cm) and cleaver (27 cm×10 cm). The ground truth pictures and images generated by AIM are shown in Figure 21(a), (b), (d), and (e). We see that our images clearly show the outline of the weapons, and the similarity metrics for the gun and cleaver are 0.81 and 0.87, respectively, as shown in Table 2.

Object	Setup	s	h_e	w_e	r_e
Rectangle	LoS	0.89	0.02	0.08	0.07
Triangle	LoS	0.85	0.03	0.07	0.10
Diamond	LoS	0.86	0.06	0.03	0.03
Circle	LoS	0.79	0.10	0.01	0.11
Hollow rect.	LoS	0.72	0.06	0.07	0.01
Gun	LoS	0.81	0.01	0.02	0.03
Cleaver	LoS	0.87	0.04	0.03	0.07
Gun	In bag	0.76	0.05	0.02	0.06
Cleaver	In bag	0.75	0.02	0.13	0.14
Gun	Under clothes	0.76	0.02	0.02	0.00
Cleaver	Under clothes	0.77	0.08	0.06	0.14
Hollow rect.	Under music	0.70	0.13	0.11	0.02
Hollow rect.	Under voice	0.71	0.07	0.09	0.02
Rectangle	60 cm Dist	0.84	0.01	0.10	0.11
Rectangle	80 cm Dist	0.70	0.06	0.11	0.20
Rectangle	100 cm Dist	0.64	0.16	0.23	0.09

Table 2: Performance metrics for various experiments.

In-bag imaging: We evaluate AIM when the weapons are put in a black trash bag. Visually, we cannot see what object is in the bag. However, as shown by Figure 1, acoustic signals are able to penetrate through the bag and sense the object. As shown in Figure 21(c), (f), and Table 2, the images still capture the shapes of the weapons, but the similarity metrics for the gun and cleaver are reduced to 0.76 and 0.75, respectively. The degradation is because acoustic signals are attenuated when penetrating through the bag.

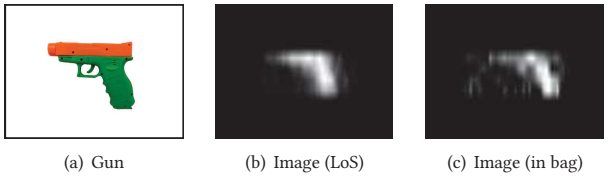
Under-clothes imaging: To explore under-clothes weapon detection, we let a person wear a hoodie with a front pocket. The weapons are put in the pocket and not visible. We sweep the mobile at 40 cm distance in front of the person and generate images using AIM. When weapons are not present, there is only noise in the image, as shown in Figure 22(a). This is because (i) the reflection from clothes is weak and (ii) body reflection is weak as the acoustic signals are damped by multiple layers of clothes on the person before arriving at his body. When the weapons are present, acoustic signals penetrating the pocket are reflected by the weapons. These signals are received by the microphone and used to generate images, as shown in Figure 22(b) and (c). From these images, we can see the shapes of weapons. The similarity metrics for the gun and cleaver



(a) Rectangle (b) Diamond (c) Triangle (d) Circle (e) Hollow rect.

(f) Rectangle (g) Diamond (h) Triangle (i) Circle (j) Hollow rect.

Figure 20: Imaging various shapes. 1) First row: objects; 2) second row: images.



(a) Gun (b) Image (LoS) (c) Image (in bag)

(d) Cleaver (e) Image (LoS) (f) Image (in bag)

Figure 21: Imaging weapons.

(a) No weapon (b) Gun (c) Cleaver

Figure 22: Imaging weapons under clothes.

(a) Music (b) Human voice

Figure 23: Imaging under environment noise.

are 0.76 and 0.77, respectively. If the weapons are hidden under multiple layers of clothes, speakers with high volume are required.

Imaging under environment noise: We evaluate the imaging performance under environment noise. We consider 1) different types of music (Jazz, Pop, and Classic) played together using the same volume as imaging signals and 2) two people keep talking during the experiment. The noise sources are 1 m away from the mobile. Under these conditions, the imaging results for a hollow rectangle are shown in Figure 23 and Table 2. Compared to the case without noise, the performance does not degrade, because the environment noise is usually lower than 10 KHz [62], while our signals for imaging are above 10 KHz.

Impact of nearby objects: To evaluate the impact of background reflection from objects close to the target, we place a whiteboard near our experiment setup. The board is in the yz-plane as shown by Figure 4(a) and at 0.5 m distance from the center of synthetic aperture. In this way, the board does not occlude the target reflection but introduces background reflection with one-way propagation distance 0.4 - 0.6 m. Such reflection is not removed in our interference cancellation, because it has similar propagation distance as the target reflection. To clearly observe its impact, we compare the images of a bar with and without the board, as shown in Figure 24(a) and (b). We see that background interference introduces artifacts in the image. To quantify them, we count the pixels with magnitude larger than 0.1 (barely visible) outside the target area. Without the board, the image only has 1% such pixels, which indicates that the image is clean. With the board, the percentage increases to 7%. When the board is moved 0.1 m away from the target, the amount of artifacts is reduced to 3%, as shown in Figure 24(c). In this case, background interference is partially separable from the target reflection, and can be mitigated by our interference cancellation. Also, the interference strength is reduced due to increased propagation distance.

Imaging at various distances: To evaluate imaging performance at a longer distance, we produce the images for a rectangle at 0.6 m, 0.8 m, 1.0 m from the mobile. The results are shown in Figure 25. We observe that the image quality degrades as the imaging distance increases due to the reduced received signal strength. At 0.6 m, the shape of the target is preserved and the similarity between the image and ground truth is 0.84, as shown in Table 2. At 0.8 m and 1 m, the similarity values reduce to 0.7 and 0.64, respectively.

9 RELATED WORK

Acoustic based imaging and sensing: Holography is a commonly used approach in acoustic imaging [23, 26, 36]. It uses 2-D receiver array to collect signals reflected by the target object, and applies 2-D FFT on the received signals for imaging. If a 2-D array is not available, a single receiver with precisely controlled movement is used to emulate it.

(a) No board (b) Board at 0.5 m (c) Board at 0.6 m

Figure 24: Imaging under background interference.

(a) 60 cm (b) 80 cm (c) 100 cm

Figure 25: Imaging at various distances.

Pulse-echo based approach is also widely used [13, 21]. It transmits acoustic pulses and estimates the propagation delay of received signals, which is used to locate a point on the object contour. By scanning over the space, the object contour can be constructed. Based on this approach, harmonic imaging [58] and multi-beam imaging [37, 48] are developed to further improve the image quality. Similar to holography, pulse-echo based approach requires a transceiver array or scannable transceiver.

SOund Navigation And Ranging (*i.e.*, *SONAR*) [21, 24, 35] is used to detect and image objects with sounds, typically for long-distance underwater scenarios. These systems are based on pulse-echo methods or synthetic aperture techniques. Directional receivers and receiver arrays are commonly used to determine the angle of arrival and increase the detection range [24].

Ultrasound imaging systems are developed using pulse-echo based methods, and widely used for medical purpose [29] and weapon detection [2]. Due to the low propagation speed and wide bandwidth of ultrasound, the delay of an echo can be accurately measured. Ultrasonic transducer arrays or scannable transducers are required to construct an image.

Recently, several new acoustic imaging systems are developed. In [42], the authors design a system to detect the object edges using acoustic meta-materials. The system achieves 1 cm resolution, but meta-materials are not widely available, which limits its applications. In [27, 53], the authors develop acoustic imaging systems based on planar microphone arrays. Wideband beamforming [27] and adaptive beamforming [53] are used to generate images. To achieve good image resolution, both systems use sizable microphone arrays (about 20 cm×20 cm), which is unacceptable for smartphones. [37] fuses measurements of a microphone array and a camera to generate 3-D surface geometry of an object. The systems developed in [32] and [3] augment FPGA platforms and smartphones with external ultrasound transducer arrays for imaging. In [6], the authors develop a system to image a mannequin around the corner using SAR techniques. To avoid trajectory errors, a motion controller is used to precisely control the movement of the speaker and microphone.

Different from existing work, AIM is the first acoustic imaging system using only an off-the-shelf mobile. For imaging, we move a mobile by hand to mimic a microphone array, and develop MPGA to remove the impact of hand jitters. Since speakers and microphones on mobiles are omni-directional, we develop a 2-stage cancellation scheme to minimize the self and background interference. Furthermore, smartphone speakers and microphones are not optimized for imaging and introduce severe distortion, so we develop a focus and denoise algorithm to remove the impact of distortion.

Acoustic signals are also used to measure the distance between the mobile and target [40, 41, 46, 47, 61, 64], which can be combined with our approach for simultaneous ranging and imaging.

RF based imaging: RF based imaging radars are widely deployed [11, 14, 18, 51, 52], which transmit RF signals up to GHz with high power [43]. Different signals have been adopted, including pulses, modulated, and unmodulated continuous waves. Synthetic aperture technique is developed for imaging radars [12, 38, 43, 43] to improve resolution and coverage. Our system also relies on SAR technique but faces new challenges due to low carrier frequency, short target distance, slow propagation of the sound, and limited hardware.

[65] develops an imaging system with 60 GHz transceivers. It uses RSS series analysis to determine 1-D information of target objects (*e.g.*, height and width). Our approach uses both magnitude and phase to obtain 2-D images and requires a simpler setup (*i.e.*, sweeping a mobile). [31] and [60] leverage RSS attenuation when signals pass through an object to image its cross section. [31] uses WiFi signals, while [60] is based on RFID. To set up these systems, [31] needs to use two drones to transmit and receive WiFi signals, while [60] requires deploying two RFID arrays at favored locations. In comparison, our approach only needs a smartphone and requires minimum setup, which is critical for ubiquitous imaging. [1] captures human skeletons through a wall with an antenna array sending signals over 5.46 − 7.24 GHz. This approach is tailored for moving human figure, whereas we can image general objects.

Interference cancellation: Several RF interference cancellation schemes are developed [8, 9]. The major contribution comes from RF and analog cancellation, which cannot be applied to our case. [45] develops secure acoustic communication by sending and canceling the jamming signals. Its interference cancellation is similar to Stage 1 in our 2-stage scheme but does not consider the AGC scaling. Our approach achieves 6 dB higher cancellation by considering AGC, and additional 5 dB cancellation using Stage 2.

10 CONCLUSION

We develop a smartphone based acoustic imaging system. Our innovation consists of (i) a new phase error correction algorithm for imaging *close-by* objects using *slowly* propagating acoustic signals with *low* carrier frequency through *digital processing*, and (ii) an algorithm to remove the blur caused by the speaker and microphone distortion. Our implementation and experiments show that it is feasible to image an object under LoS, under-clothes, and in-bag scenarios. As part of future work, we will improve the image quality at longer distance and extend our system to 3-D imaging.

11 ACKNOWLEDGEMENTS

This work is supported in part by NSF Grant CNS-1718585. We are grateful to Fadel Adib, Preston S. Wilson, and anonymous reviewers for their insightful comments and help.

REFERENCES

[1] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Fredo Durand. 2015. Capturing the Human Figure Through a Wall. In *Proc. of SIGGRAPH*.

[2] Alan Agurto, Yong Li, Gui Yun Tian, Nick Bowring, and Stephen Lockwood. 2007. A review of concealed weapon detection and research in perspective. In *Networking, Sensing and Control, 2007 IEEE International Conference on*. IEEE, 443–448.

[3] Sewoong Ahn, Jeeun Kang, Pilsu Kim, Gunho Lee, Eunji Jeong, Woojin Jung, Minsuk Park, and Tai-kyong Song. 2015. Smartphone-based portable ultrasound imaging system: Prototype implementation and evaluation. In *Ultrasonics Symposium (IUS), 2015 IEEE International*. IEEE, 1–4.

[4] Abc Apps. 2017. Sound Meter. (2017). https://play.google.com/store/apps/details?id=com.gamebasic.decibel&hl=en.

[5] Audio-Technica. 2017. Audio-Technica ATH-A1000Z Art Series Headphone. (2017). https://audiocubes.com/collections/headphones/products/audio-technica-ath-a1000z-art-series-headphone.

[6] Hisham Bedri, Micha Feigin, Michael Everett, Gregory L Charvat, Ramesh Raskar, et al. 2014. Seeing around corners with a mobile phone?: synthetic aperture audio imaging. In *ACM SIGGRAPH 2014 Posters*. ACM, 84.

[7] Fabrizio Berizzi, Marco Martorella, Brett Haywood, Enzo Dalle Mese, and Silvia Bruscoli. 2004. A survey on ISAR autofocusing techniques. In *Image Processing, 2004. ICIP'04. 2004 International Conference on*, Vol. 1. IEEE, 9–12.

[8] Dinesh Bharadia and Sachin Katti. 2014. Full duplex MIMO radios. *Self* 1, A2 (2014), A3.

[9] Dinesh Bharadia, Emily McMilin, and Sachin Katti. 2013. Full duplex radios. *ACM SIGCOMM Computer Communication Review* 43, 4 (2013), 375–386.

[10] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[11] Graham Brooker. 2009. *Introduction to sensors for ranging and imaging*. The Institution of Engineering and Technology.

[12] W.G. Carrara, R.S. Goodman, and R.M. Majewski. 1995. *Spotlight Synthetic Aperture Radar: Signal Processing Algorithms*. Artech House.

[13] Vincent Chan and Anahi Perlas. 2011. Basics of ultrasound imaging. In *Atlas of ultrasound-guided procedures in interventional pain management*. Springer, 13–19.

[14] Margaret Cheney and Brett Borden. 2009. *Fundamentals of radar imaging*. SIAM.

[15] Microsoft Corporation. 2011. Microsoft LifeCam Studio. (2011). https://www.microsoft.com/accessories/en-us/products/webcams/lifecam-studio/q2f-00013.

[16] Android Developers. 2017. Automatic Gain Control. (2017). https://developer.android.com/reference/android/media/audiofx/AutomaticGainControl.html.

[17] PH Eichel and CV Jakowatz. 1989. Phase-gradient algorithm as an optimal estimator of the phase derivative. *Optics letters* 14, 20 (1989), 1101–1103.

[18] Charles Elachi. 1988. Spaceborne radar remote sensing: applications and techniques. *New York, IEEE Press, 1988, 285 p.* (1988).

[19] Joachim HG Ender and Jens Klare. 2009. System architectures and algorithms for radar imaging by MIMO-SAR. In *Radar Conference, 2009 IEEE*. IEEE, 1–6.

[20] Matteo Frigo and Steven G. Johnson. 2005. The Design and Implementation of FFTW3. *Proc. IEEE* 93, 2 (2005), 216–231. Special issue on "Program Generation, Optimization, and Platform Adaptation".

[21] Woon Siong Gan. 2012. *Acoustical Imaging: Techniques and Applications for Engineers*. John Wiley & Sons.

[22] Hirad Ghaemi, Michele Galletti, Thomas Boerner, Frank Gekat, and Mats Viberg. 2009. CLEAN technique in strip-map SAR for high-quality imaging. In *Aerospace conference, 2009 IEEE*. IEEE, 1–7.

[23] Jorgen Hald. 2001. Time domain acoustical holography and its applications. *Sound and Vibration* 35, 2 (2001), 16–25.

[24] Roy Edgar Hansen. 2011. *Introduction to synthetic aperture sonar*. INTECH Open Access Publisher.

[25] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. 2015. *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

[26] Bernard Percy Hildebrand. 2013. *An introduction to acoustical holography*. Springer Science & Business Media.

[27] Alberto Izquierdo, Juan José Villacorta, Lara del Val Puente, and Luis Suárez. 2016. Design and Evaluation of a Scalable and Reconfigurable Multi-Platform System for Acoustic Imaging. *Sensors* 16, 10 (2016), 1671.

[28] Charles VJ Jakowatz, Daniel E Wahl, Paul H Eichel, Dennis C Ghiglia, and Paul A Thompson. 2012. *Spotlight-Mode Synthetic Aperture Radar: A Signal Processing Approach*. Springer Science & Business Media.

[29] Jørgen Arendt Jensen. 2007. Medical ultrasound imaging. *Progress in biophysics and molecular biology* 93, 1 (2007), 153–165.

[30] Peter K. 2015. Loud, louder, LOUDEST! Here are the phones with the loudest speakers so far in 2015. (2015). https://www.phonearena.com/news/Loud-louder-LOUDEST-Here-are-the-phones-with-the-loudest-speakers-so-far-in-2015_id74023.

[31] Chitra R Karanam and Yasamin Mostofi. 2017. 3D through-wall imaging with unmanned aerial vehicles using wifi. In *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks*. ACM, 131–142.

[32] Gi-Duck Kim, Changhan Yoon, Sang-Bum Kye, Youngbae Lee, Jeeun Kang, Yangmo Yoo, and Tai-Kyong Song. 2012. A single FPGA-based portable ultrasound imaging system for point-of-care applications. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control* 59, 7 (2012), 1386–1394.

[33] Voon Chet Koo, Tien Sze Lim, and Hean Teik Chuah. 2005. A comparison of autofocus algorithms for SAR imagery. In *Progress in electromagnetics research symposium*, Vol. 1. 16–19.

[34] Krzysztof Kulpa. 2008. The CLEAN type algorithms for radar signal processing. In *Microwaves, Radar and Remote Sensing Symposium, 2008. MRRS 2008*. IEEE, 152–157.

[35] L Jay Larsen, Andy Wilby, and Colin Stewart. 2010. Deep ocean survey and search using synthetic aperture sonar. In *OCEANS 2010*. IEEE, 1–4.

[36] Hua Lee. 2016. *Acoustical Sensing and Imaging*. CRC Press.

[37] Mathew Legg and Stuart Bradley. 2014. Automatic 3D scanning surface generation for microphone array acoustic imaging. *Applied Acoustics* 76 (2014), 230–237.

[38] Bassem R Mahafza. 2002. *Radar systems analysis and design using MATLAB*. CRC press.

[39] Wenguang Mao. 2017. Proof of Theorem 1. (2017). http://www.cs.utexas.edu/~wmao/ReferenceLink/mobisys18theorem1.pdf.

[40] Wenguang Mao, Jian He, and Lili Qiu. 2016. CAT: High-Precision Acoustic Motion Tracking. In *Proc. of ACM MobiCom*.

[41] Wenguang Mao, Zaiwei Zhang, Lili Qiu, Jian He, Yuchen Cui, and Sangki Yun. 2017. Indoor Follow Me Drone. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 345–358.

[42] Miguel Molerón and Chiara Daraio. 2015. Acoustic metamaterial for subwavelength edge detection. *Nature communications* 6 (2015).

[43] Alberto Moreira, Pau Prats-Iraola, Marwan Younis, Gerhard Krieger, Irena Hajnsek, and Konstantinos P Papathanassiou. 2013. A tutorial on synthetic aperture radar. *IEEE Geoscience and Remote Sensing Magazine* 1, 1 (2013), 6–43.

[44] Alberto Moreira, Pau Prats-Iraola, Marwan Younis, Gerhard Krieger, Irena Hajnsek, and Konstantinos P Papathanassiou. 2013. A tutorial on synthetic aperture radar. *IEEE Geoscience and remote sensing magazine* 1, 1 (2013), 6–43.

[45] Rajalakshmi Nandakumar, Krishna Kant Chintalapudi, and Venkata N. Padmanabhan. 2013. Dhwani : Secure Peer-to-Peer Acoustic NFC. In *Proc. of ACM SIGCOMM*.

[46] Rajalakshmi Nandakumar, Shyam Gollakota, and Nathaniel Watson. 2015. Contactless Sleep Apnea Detection on Smartphones. In *Proc. of ACM MobiSys*.

[47] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. 2016. FingerIO: Using Active Sonar for Fine-Grained Finger Tracking. In *Proc. of ACM CHI*. 1515–1525.

[48] Clement Papadacci, Mathieu Pernot, Mathieu Couade, Mathias Fink, and Mickaël Tanter. 2014. High-contrast ultrafast imaging of the heart. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control* 61, 2 (2014), 288–301.

[49] J Piechowicz. 2011. Sound Wave Diffraction at the Edge of a Sound Barrier. *Acta Physica Polonica, A.* 119 (2011).

[50] Markku Pukkila. 2000. Channel estimation modeling. *Nokia Research Center* (2000).

[51] Andreas Reigber, Rolf Scheiber, Marc Jager, Pau Prats-Iraola, Irena Hajnsek, Thomas Jagdhuber, Konstantinos P Papathanassiou, Matteo Nannini, Esteban Aguilera, Stefan Baumgartner, et al. 2013. Very-high-resolution airborne synthetic aperture radar imaging: Signal processing and applications. *Proc. IEEE* 101, 3 (2013), 759–783.

[52] Mark A Richards. 2005. *Fundamentals of radar signal processing*. Tata McGraw-Hill Education.

[53] Feng Su and Chris Joslin. 2015. Acoustic imaging using a 64-node microphone array and beamformer system. In *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 168–173.

[54] GSMArena Team. 2016. Samsung Galaxy S7 vs. Apple iPhone 6s: Loudspeaker, audio quality. (2016). https://www.gsmarena.com/galaxy_s7_vs_iphone_6s-review-1415p5.php.

[55] Douglas G Thompson, James S Bates, and David V Arnold. 1999. Extending the phase gradient autofocus algorithm for low-altitude stripmap mode SAR. In *Radar Conference, 1999. The Record of the 1999 IEEE*. IEEE, 36–40.

[56] New York Times. 2017. Gunman Kills at Least 26 in Attack on Rural Texas Church. (2017). https://www.nytimes.com/2017/11/05/us/church-shooting-texas.html.

[57] David Tse and Pramod Viswanath. 2005. *Fundamentals of wireless communication*. Cambridge university press.

[58] Paul LMJ van Neer, Mikhail G Danilouchkine, Martin D Verweij, Libertario Demi, Marco M Voormolen, Anton FW van der Steen, and Nico de Jong. 2011. Comparison of fundamental, second harmonic, and superharmonic imaging: A simulation study. *The Journal of the Acoustical Society of America* 130, 5 (2011), 3148–3157.

[59] DE Wahl, PH Eichel, DC Ghiglia, and CV Jakowatz. 1994. Phase gradient autofocus-a robust tool for high resolution SAR phase correction. *IEEE Trans. Aerospace Electron. Systems* 30, 3 (1994), 827–835.

[60] Ju Wang, Jie Xiong, Xiaojiang Chen, Hongbo Jiang, Rajesh Krishna Balan, and Dingyi Fang. 2017. TagScan: Simultaneous target imaging and material identification with commodity RFID devices. In *Proc. ACM MobiCom*. 1–14.

[61] Wei Wang, Alex X Liu, and Ke Sun. 2016. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, 82–94.
[62] Inc Waterline Media. 2003. The Frequency Spectrum, Instrument Ranges, and EQ Tips. (2003). http://www.guitarbuilding.org/wp-content/uploads/2014/06/Instrument-Sound-EQ-Chart.pdf.
[63] Wikipedia. 2018. 2017 Las Vegas shooting. (2018). https://en.wikipedia.org/wiki/2017_Las_Vegas_shooting.
[64] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-Grained Acoustic-based Device-Free Tracking. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 15–28.
[65] Yanzi Zhu, Yibo Zhu, Ben Y. Zhao, and Haitao Zheng. 2015. Reusing 60GHz Radios for Mobile Radar Imaging. In *Proc. of ACM MobiCom*.