Coherent 3D Acoustic Imaging on a Smartphone

Aryan Mahindra

March 2026

Abstract

We present a prototype for coherent 3D acoustic imaging on a smartphone. Using an unmodified iPhone 15 Pro, the system emits 18–22 kHz FMCW chirps, records echoes on the built-in microphone, and reconstructs a 3D acoustic reflectivity volume entirely on-device with GPU backprojection. A motorized turntable rotates the object, inducing a known circular synthetic aperture and removing the need for freehand trajectory estimation. The current prototype operates at a 0.30 m standoff and completes on-device processing in under 30 s. At the current 4 kHz bandwidth, the nominal range resolution is 4.3 cm; under ideal full-aperture integration, the nominal cross-range scale is 8.6 mm ($\lambda/2$ at 20 kHz). We show qualitative reconstructions of two cylindrical test objects, an aluminum can and a stainless-steel tumbler, and recover the main reflective structure of each. We are not aware of prior published smartphone systems that combine built-in acoustics with coherent 3D volumetric reconstruction without object-specific priors. The main gaps in the current evaluation are quantitative accuracy metrics, broader object coverage, and coherence ablations.

1 Introduction

Prior phone acoustic systems reach sparse tracking, 2D imaging, or constrained 3D outputs, but not coherent 3D volumetric reconstruction of nearby general objects. FingerIO, LLAP, MilliSonic, and ReflecTrack track sparse points [6, 13, 11, 15]. AIM [5] synthesizes a 1D aperture from linear phone motion and reconstructs a 2D acoustic image. SONDAR [4] adapts inverse synthetic aperture ideas to commodity devices for 2D size and shape measurement. SonicHand [12] produces a 3D hand output, but only under a fixed-topology learned prior. Laboratory SAS systems achieve coherent 3D reconstruction, but rely on specialized hardware and controlled acquisition [1, 8, 9].

We treat the setup as an acoustic inverse synthetic aperture imaging problem. A stationary iPhone observes a rotating target; in the object frame, this is equivalent to a virtual circular aperture. This removes freehand trajectory estimation, a major source of error in prior phone-based SAR [5], by making the acquisition geometry mechanically controlled. All acoustic sensing and reconstruction run on the phone; the turntable provides rotation only.

This paper presents the system formulation, a working prototype, and initial qualitative results. We report the operating regime, theoretical resolution limits, and on-device runtime. The current evidence is qualitative and should be read as feasibility, not validation.

1.1 Contributions

1. A coherent 3D acoustic imaging system that uses the iPhone's built-in speaker and microphone, with a motorized turntable providing controlled object rotation.

2. An acoustic ISAR formulation in which target rotation induces a known circular synthetic aperture, eliminating freehand trajectory estimation.

3. An end-to-end on-device pipeline comprising FMCW acquisition, phase alignment, GPU backprojection, and point-cloud extraction on an iPhone 15 Pro.

2 Related work and scope of novelty

2.1 Phone acoustics: tracking, mapping, and 2D imaging

Much prior work in smartphone acoustics focuses on estimating only a small number of geometric degrees of freedom. FingerIO [6] tracks a fingertip in 2D using active sonar. LLAP [13] achieves millimeter-scale displacement tracking using phase changes in acoustic signals. MilliSonic [11] pushes tracking precision further, while ReflecTrack [15] estimates 3D position using a dual-microphone smartphone and environmental reflections. These systems produce trajectories or sparse point sets rather than spatial images.

Two strands of prior work move smartphone acoustics closer to imaging. AIM [5] uses linear sensor motion to synthesize an aperture and reconstruct 2D acoustic images. Li et al. [3] reconstruct a 2D acoustic heat map from swept-phone measurements. SONDAR [4] adapts inverse synthetic aperture ideas to commodity devices for 2D size and shape measurement. BatMapper [14] and SAMS [7] reconstruct room-scale structure, not nearby tabletop objects.

2.2 3D outputs under different assumptions

SonicHand [12] reconstructs a 3D hand skeleton from smartphone audio, but its output is a fixed set of hand joints under a strong task-specific prior. Shih and Rowe [10] reconstruct room geometry, but use a separate fixed speaker module. At laboratory scale, in-air synthetic-aperture sonar systems and neural volumetric methods achieve coherent 3D reconstruction with specialized hardware and off-phone compute [1, 8]. Classic acoustic ISAR has also been demonstrated with research sonar hardware [9].

2.3 Scope of novelty

Table 1 summarizes this comparison. The closest published systems satisfy at most two of the three properties we care about: phone-native acoustics, coherent 3D volumetric output, and no object-specific prior. Based on literature available through March 2026, we are not aware of prior published smartphone systems that combine built-in acoustics with coherent 3D volumetric reconstruction in a single on-device pipeline without object-specific priors. The current evaluation is limited to two convex reflective objects. The formulation is object-general, but the present experiments are not.

3 System overview

Figure 1 summarizes the operating principle. The iPhone is placed at a fixed standoff from a motorized turntable. The phone emits near-ultrasonic FMCW chirps and records the resulting echoes. In the object frame, the stationary phone becomes a virtual circular aperture around the target (Figure 2). We reconstruct the scene by coherent bistatic backprojection.

Table 1: Prior-art comparison for smartphone acoustic reconstruction. “Phone-native acoustics” means all acoustic transmit/receive uses only built-in phone transducers. Our system uses a motorized turntable for controlled object rotation but no external acoustic sensor, array, or compute server.

System	Phone acoustics	3D volume	No prior	Notes
AIM [5]	✓	—	✓	2D acoustic image from linear phone motion.
Li et al. [3]	✓	—	✓	2D acoustic heat map from swept phone.
SONDAR [4]	✓	—	✓	ISAR-style 2D image for shape/size.
SonicHand [12]	✓	joints only	—	Fixed hand skeleton, not general volume.
Shih & Rowe [10]	—	✓	—	Room surfaces; separate speaker module.
AirSAS + Neural SAS [1, 8]	—	✓	✓	Lab SAS with specialized hardware.
Ours	✓[*]	✓	✓[†]	Coherent 3D reflectivity volume; turntable supplies rotation only.

[*] Motorized turntable for object rotation; no external acoustic hardware.

[†] No object-specific prior in formulation; evaluated on two convex reflective objects to date.

3.1 Acquisition

The prototype transmits a linear chirp from 18 kHz to 22 kHz ($B = 4\,\text{kHz}$, $T \approx 10\,\text{ms}$) at a 48 kHz sample rate. In our setup, usable output is limited to roughly a 4 kHz band: below 18 kHz the chirp becomes more audible, and above 22 kHz output falls off rapidly. We configure the iOS audio session in measurement mode and disable automatic gain control, echo cancellation, and related speech processing to preserve phase and amplitude. Frames are acquired at 13.3 Hz, set by the chirp-plus-guard cycle. At a turntable period of 15 s/rev, this yields roughly 200 aspect samples per turn. We use 200-, 400-, and 600-frame scan settings.

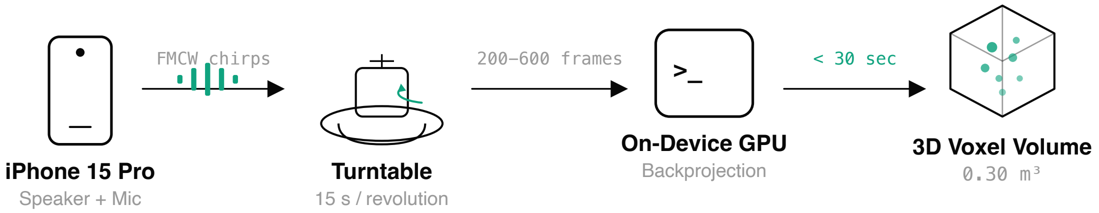
3.2 Calibration

Each scan begins with an empty-turntable calibration. We average an empty-turntable scan to estimate static background and direct speaker-to-microphone leakage. We estimate standoff and turntable center by fitting the dominant reflector radius across angle-indexed frames. Backprojection is sensitive to center and standoff errors because both change the predicted path lengths.

3.3 Signal processing pipeline

The processing pipeline has three stages, shown in Figure 3.

(1) Frame formation. Each received echo is range-compressed against the transmitted chirp replica. The recorded waveform is multiplied by the conjugate of the transmit chirp, producing a beat signal whose frequency is proportional to round-trip delay ($f_{\text{beat}} = \beta\tau$, where $\beta = B/T$ and τ is the path delay). A windowed FFT of the beat signal yields a complex range profile, $b_k(\rho)$. This range profile preserves complex phase, which is later compensated at the center frequency during coherent backprojection.



End-to-end pipeline: acquisition through reconstruction on a single unmodified smartphone

Figure 1: System overview. A stationary iPhone emits near-ultrasonic FMCW chirps while a motorized turntable rotates the target. In the target frame, the fixed phone becomes an equivalent circular synthetic aperture, producing angle-indexed complex range profiles that are coherently backprojected into a 3D voxel volume on-device.

(2) Coherence maintenance. Background subtraction removes the static direct path and clutter measured during empty-turntable calibration. We phase-align each frame to a common reference by multiplying frame k by $\exp(-j\varphi_k)$, where φ_k is the phase of the strongest near-zero-range component (the residual direct path after background subtraction). This compensates for frame-to-frame phase offset. A per-frame consistency check discards frames whose reference phase deviates by more than $\pi/4$ from the running estimate. This threshold was chosen empirically from calibration runs. Without phase normalization, the coherent reconstruction visibly defocuses and begins to resemble incoherent accumulation.

(3) 3D reconstruction. Given the turntable angle for each frame, we synthesize acquisition poses in the object frame and coherently backproject the complex profiles into a 3D voxel grid. We evaluate range profiles at non-grid path lengths using linear interpolation. After accumulation, we threshold the volume and apply 3D non-maximum suppression to extract isolated local maxima as the output point cloud.

4 Acoustic ISAR formulation

4.1 Transmit and receive model

We write the transmitted FMCW chirp as

$$s_{\mathrm{tx}}(t) = \exp\Big(j2\pi\Big(f_0 t + \tfrac{1}{2}\beta t^2\Big)\Big), \quad 0 \le t \le T, \tag{1}$$

where $f_0 = 18\,\mathrm{kHz}$ is the start frequency, $\beta = B/T$ is the chirp rate, and $B = 4\,\mathrm{kHz}$ is the swept bandwidth. For voxel $\mathbf{v} \in \mathbb{R}^3$ and frame k, the bistatic path length is

$$d_k(\mathbf{v}) = \|\mathbf{v} - \mathbf{t}_k\| + \|\mathbf{v} - \mathbf{r}_k\|, \tag{2}$$

where $\mathbf{t}_k$ and $\mathbf{r}_k$ are the speaker and microphone positions expressed in the object frame. On the iPhone 15 Pro, the speaker and microphone centers are approximately 16 mm apart; at 0.30 m standoff, the bistatic

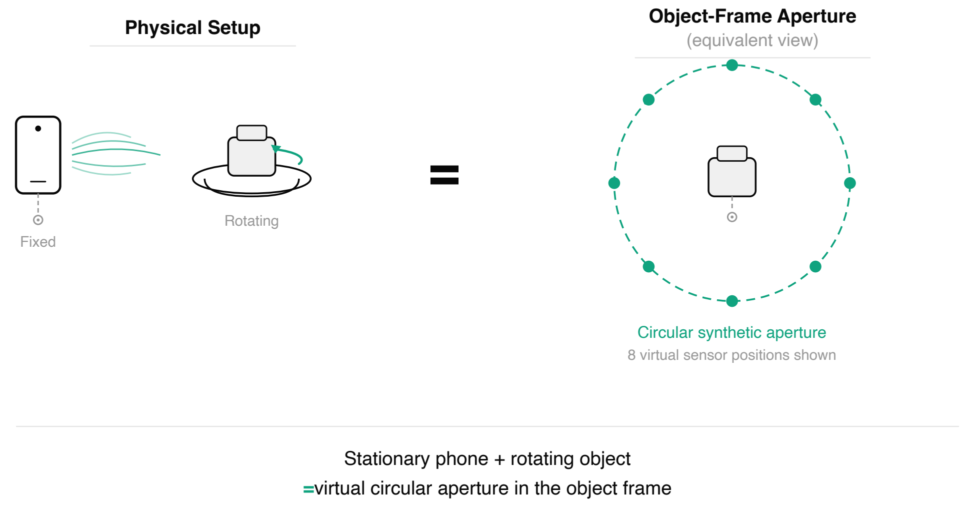

Physical Setup
Fixed
Rotating
=
Object-Frame Aperture
(equivalent view)
Circular synthetic aperture
8 virtual sensor positions shown
Stationary phone + rotating object
=virtual circular aperture in the object frame

Figure 2: ISAR equivalence. Left: physical setup with a fixed phone and a rotating object. Right: equivalent circular synthetic aperture in the object frame.

angle is roughly 3°. We retain the full bistatic path length in Eq. (2) for phase accuracy. The corresponding delay is $\tau_k(\mathbf{v}) = d_k(\mathbf{v})/c$, with c the speed of sound in air.

Although the phone is stationary, the object rotates. If the turntable angle at frame k is θ_k about the vertical axis, the equivalent acquisition poses in the object frame are

$$\mathbf{t}_k = \mathbf{R}_z(-\theta_k)\,\mathbf{t}_0, \qquad \mathbf{r}_k = \mathbf{R}_z(-\theta_k)\,\mathbf{r}_0, \tag{3}$$

where $\mathbf{R}_z(-\theta_k)$ denotes rotation about the turntable's vertical axis, and $\mathbf{t}_0$ and $\mathbf{r}_0$ are the fixed phone transducer locations. This turns object rotation into a known synthetic aperture.

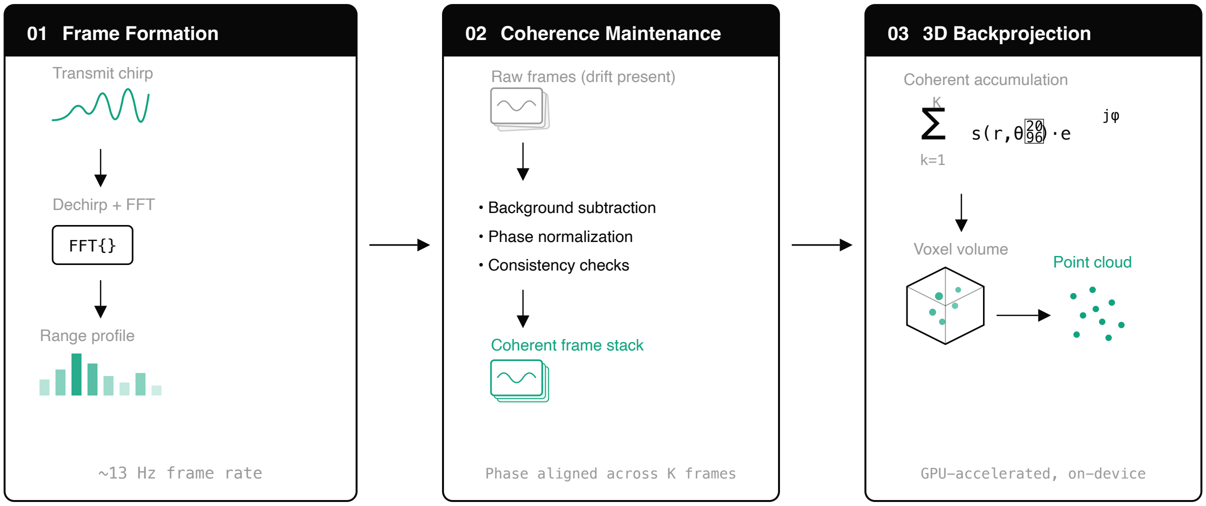
4.2 Coherent backprojection

After range compression, frame k yields a complex range profile $b_k(\rho)$ indexed by path length ρ. We compute voxel intensity by coherent accumulation:

$$V(\mathbf{v}) = \sum_{k=1}^{K} b_k\big(d_k(\mathbf{v})\big)\exp\!\Big(-j\frac{2\pi}{\lambda_c}\,d_k(\mathbf{v})\Big), \tag{4}$$

where $\lambda_c = c/f_c$ is the center-wavelength ($f_c = 20\,\text{kHz}$, $\lambda_c \approx 17\,\text{mm}$). We define the final voxel intensity as $I(\mathbf{v}) = |V(\mathbf{v})|$. We extract a point cloud by thresholding $I(\mathbf{v})$ and retaining spatially isolated maxima.

If the geometry is correct and the frames remain coherent, returns from a true scatterer add constructively at the correct voxel. If the geometry or phase is wrong, these contributions cancel and the reconstruction defocuses.



Complete signal processing chain: raw acoustic echoes to 3D volumetric reconstruction

Figure 3: Signal processing pipeline. Stage 1: each chirp is dechirped and converted to a complex range profile. Stage 2: background subtraction, phase normalization, and consistency checks maintain coherence across the frame sequence. Stage 3: GPU-accelerated backprojection coherently accumulates the profiles into a 3D voxel volume, from which a point cloud is extracted.

4.3 Why ISAR over freehand phone SAR

In phone SAR, as in AIM [5], the same acoustic measurements must support both trajectory estimation and imaging. Freehand SAR couples these two tasks in the same noisy data. In our turntable-based ISAR setup, the acquisition geometry is mechanically determined: the phone remains fixed, the object rotates through known angular increments, and the synthetic aperture is induced rather than inferred.

5 Implementation and prototype parameters

Table 2 summarizes the prototype operating regime. The current implementation runs on an unmodified iPhone 15 Pro. Reconstruction runs on-device on the phone GPU. The reconstruction grid is $60 \times 60 \times 60$ voxels with 5 mm spacing, covering a $0.30 \times 0.30 \times 0.30$ m volume centered on the turntable. The 5 mm voxel spacing oversamples the expected acoustic point-spread function and should not be interpreted as 5 mm physical resolution.

The dominant computation is volumetric backprojection, whose cost scales as $O(N_{\text{vox}} \cdot N_{\text{frames}})$. We use a Metal GPU kernel to evaluate Eq. (4) in parallel over the voxel grid.

6 Prototype characterization

We present a prototype and characterize the regime in which it currently operates. The resolution figures below are theoretical. Measured PSF and quantitative accuracy are not yet available.

Table 2: Prototype operating regime.

Parameter	Value
Device	iPhone 15 Pro
Acoustic hardware	Built-in speaker + microphone
Speaker–mic separation	~16 mm
Chirp band	18–22 kHz (B = 4 kHz)
Center frequency f_c	20 kHz
Chirp duration T	~10 ms
Sample rate	48 kHz
Frame rate	13.3 Hz
Turntable period	~15 s/rev
Standoff distance	~0.30 m
Frames per scan	200 / 400 / 600
Reconstruction volume	0.30 × 0.30 × 0.30 m
Voxel spacing	5 mm
External hardware	Motorized turntable only
Processing location	Entirely on-device

Table 3: Prototype characterization (theoretical).

Quantity	Value	Dominant factor
Range resolution	~4.3 cm	4 kHz bandwidth (theoretical)
Cross-range scale	~8.6 mm	$\lambda/2$ at 20 kHz (theoretical)
Voxel spacing	5 mm	Grid sampling only
On-device processing	<30 s	GPU backprojection

6.1 Theoretical resolution

The nominal bandwidth-limited range resolution is

$$\Delta r \approx \frac{c}{2B}. \tag{5}$$

With $B = 4$ kHz and $c = 343$ m/s, the nominal range resolution is 4.3 cm. This assumes ideal matched filtering; in practice, windowing and speaker nonlinearity may broaden the effective response. The output should be read as a sparse reflectivity representation, not a dense surface reconstruction.

Under ideal full-aperture coherent integration, the cross-range scale is $\delta_{\text{CR}} \approx \lambda/(2\sin(\Delta\phi/2))$, which reduces to $\lambda/2$ for a full 2π rotation. At $f_c = 20$ kHz ($\lambda \approx 17$ mm), this yields an ideal tangential scale of 8.6 mm (Figure 4). The system operates in the near field of the synthetic aperture, so the backprojection in Eq. (4) uses exact spherical-wave path lengths rather than a far-field approximation.

Elevation is less constrained because the aperture samples azimuth but not height. At the current operating point, elevation is governed by the speaker's vertical beam pattern rather than by the synthetic aperture. The resulting missing-cone effect means the reconstruction is best interpreted as primarily azimuthal.

6.2 Runtime

For 600 frames (high-quality mode), on-device processing completes in under 30 s on an iPhone 15 Pro. Processing time excludes acquisition time (15–45 s depending on scan mode).

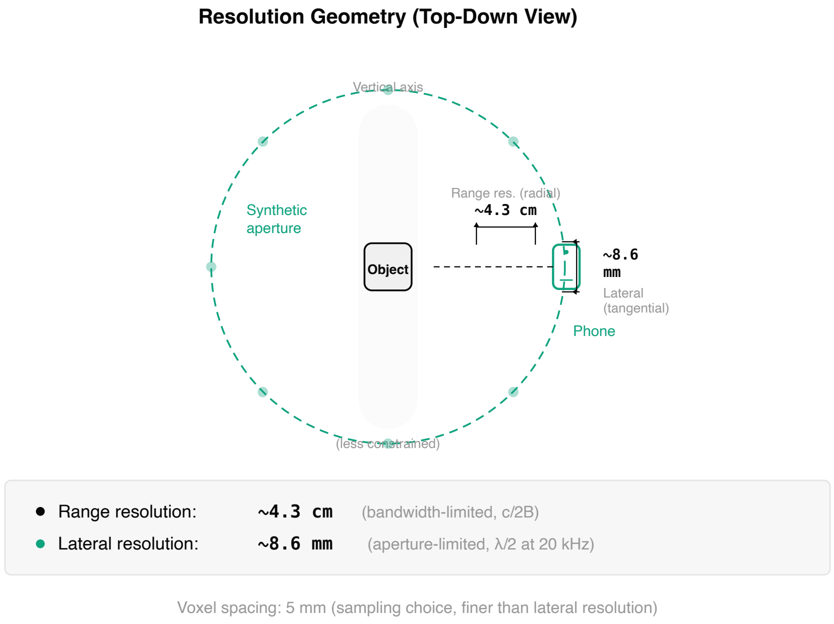

Resolution Geometry (Top-Down View)
Vertical axis
Synthetic aperture
Range res. (radial)
~4.3 cm
Object
~8.6 mm
Lateral (tangential)
Phone
(less constrained)
● Range resolution: ~4.3 cm (bandwidth-limited, c/2B)
● Lateral resolution: ~8.6 mm (aperture-limited, λ/2 at 20 kHz)
Voxel spacing: 5 mm (sampling choice, finer than lateral resolution)

Figure 4: Aperture geometry and theoretical resolution scales. The circular synthetic aperture (dashed) yields bandwidth-limited range resolution of ∼4.3 cm (radial) and aperture-limited cross-range scale of ∼8.6 mm (tangential). Shading indicates the missing-cone region.

6.3 Interpretation of the output

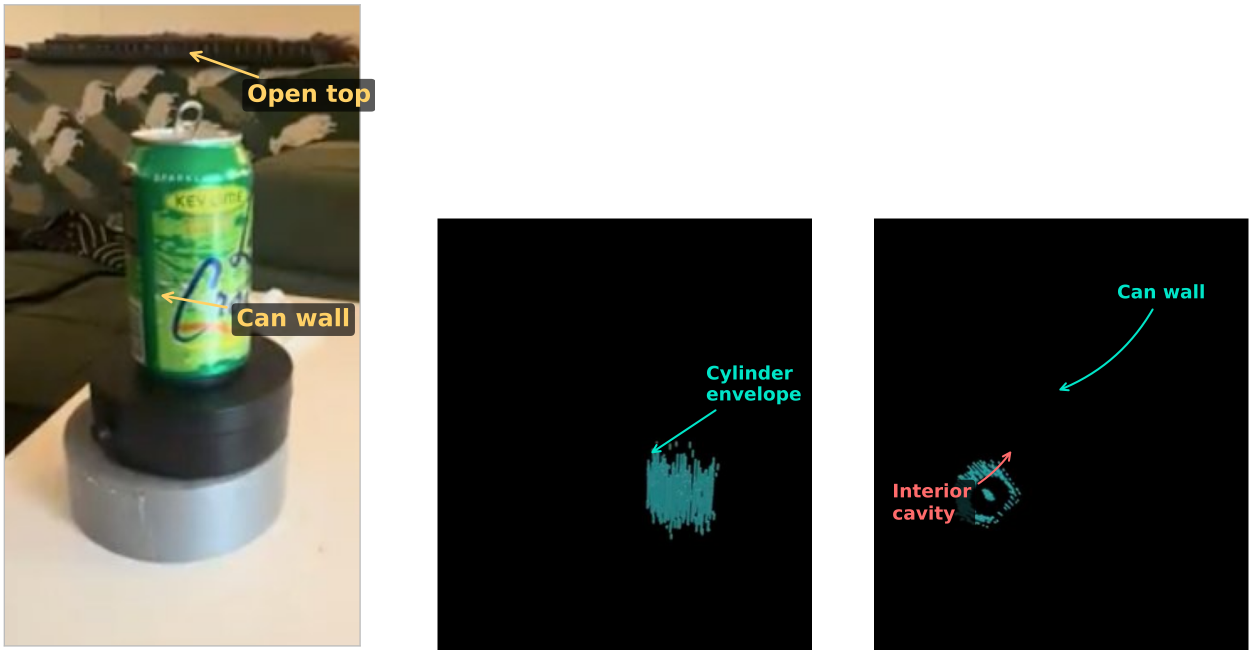
The output is a 3D reflectivity volume and a derived point cloud, not a watertight surface mesh. Strong reflectors, edges, and material discontinuities dominate the reconstruction. Weakly reflecting or unfavorably oriented surfaces may be underrepresented.

6.4 Reconstruction examples

Figures 5 and 6 show qualitative reconstruction results for two test objects: an aluminum can (66 mm diameter, 122 mm height) and a stainless-steel tumbler. Both are cylindrical and acoustically reflective; they were chosen because their known geometry simplifies visual assessment. Out of eight total scans across the two objects, six with successful calibration convergence produced similar coarse geometry across repeats; we show the median-quality successful scan for each object. The can reconstruction recovers the cylindrical wall and open rim; the tumbler reconstruction shows the body and lid region. We do not yet report quantitative accuracy metrics such as recovered-diameter error or Chamfer distance.

7 Challenges and limitations

Self-interference. The speaker and microphone are approximately 16 mm apart, so the direct path is much stronger than echoes from an object at 30 cm. We suppress direct-path leakage using empty-scene subtraction and phase-referenced frame normalization. Measured suppression ratios are not yet reported.



(a) Photograph.

(b) Side-view reconstruction.

(c) Top-view reconstruction.

Figure 5: Aluminum can (66 mm × 122 mm). (a) Photograph. (b) Side-view 3D reconstruction from the acoustic reflectivity volume. (c) Top-view reconstruction.

Phase coherence. At 20 kHz ($\lambda \approx 17$ mm), small timing errors can translate into large phase errors that defocus the reconstruction. Clock drift, thermal variation, speaker nonlinearity, and system-latency changes threaten coherence across a 15–45 s scan. Measured phase-stability diagnostics are not yet available.

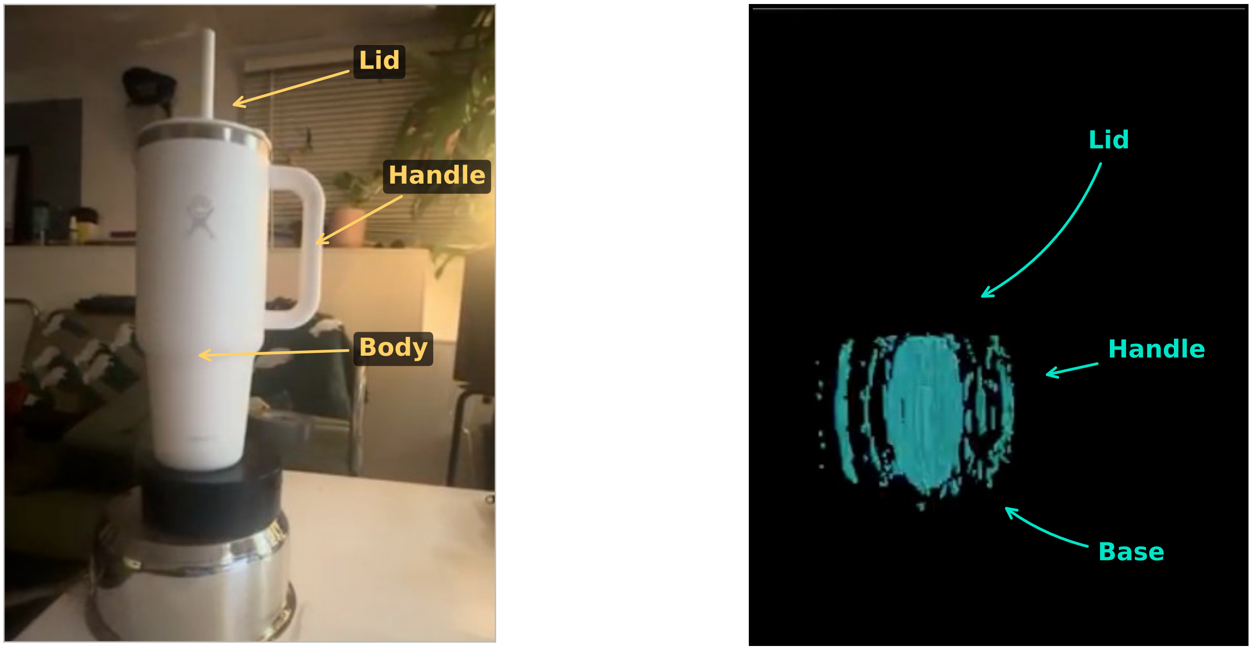
Geometry sensitivity. The ISAR pose model depends on the assumed turntable center and standoff. Even centimeter-scale errors shift predicted path lengths enough to defocus the reconstruction globally.

Bandwidth limits. The iPhone speaker provides usable energy only over a roughly 4 kHz band near 20 kHz. This limits theoretical range resolution to 4.3 cm even with ideal reconstruction.

Aperture anisotropy. A single-height circular aperture does not sample vertical spatial frequencies. The missing-cone effect means elevation is not resolved by the synthetic aperture.

Speed-of-sound variation. The reconstruction assumes $c = 343$ m/s. A 5 °C temperature change shifts c by roughly 3 m/s (0.9%), producing a range error of approximately 2.7 mm at 0.30 m standoff. No in-situ c calibration is currently performed.

Turntable angular accuracy. The current prototype derives the turntable angle θ_k from nominal motor speed rather than encoder feedback. At 20 kHz and 0.30 m standoff, a 1° angular error produces a lateral position error of roughly 5 mm ($\approx \lambda/3$), which would degrade coherent focusing.



(a) Photograph.

(b) 3D reconstruction.

Figure 6: Stainless-steel tumbler. (a) Photograph. (b) 3D acoustic reconstruction.

Controlled-setting requirement. The prototype requires the target to rotate on a turntable, restricting it to controlled tabletop scenarios.

8 Discussion

This system is not a replacement for optical 3D scanning. It shows that coherent 3D volumetric acoustic reconstruction is possible on a commodity phone when the acquisition geometry is mechanically controlled and the pipeline preserves phase. The key systems choice is to use object-centric ISAR instead of freehand SAR, replacing trajectory estimation with calibration.

The current evaluation has gaps. Quantitative accuracy metrics (Chamfer distance, recovered-diameter error) are missing. A measured point-spread function from a point-like reflector is needed. A coherent-versus-incoherent ablation would confirm that the phase pipeline is actually helping. A 2D baseline comparison (e.g., AIM-style linear SAR on the same objects) would show the value added by the 3D pipeline. The two test objects are both convex and acoustically reflective; non-convex, absorptive, or irregularly shaped targets are needed. Measured self-interference suppression ratios and phase-stability diagnostics are also needed.

PowerPhone [2] suggests that some phones can be software-reconfigured for higher sampling rates, which could improve acoustic bandwidth. Learned volumetric priors, as in recent neural SAS methods [8], could help regularize the bandwidth limitation.

9 Conclusion

We demonstrate coherent 3D acoustic reflectivity reconstruction on a phone in a controlled turntable setting. The system uses 18–22 kHz FMCW chirps, a motorized turntable, and an object-centric ISAR formulation. Reconstruction runs entirely on the iPhone 15 Pro and produces a 3D reflectivity volume and point cloud. The closest published smartphone acoustic systems either stop at 2D imaging, return sparse tracked points, or reconstruct constrained structures such as hand skeletons. Right now this is a controlled feasibility result, not a robust general-purpose scanner. Quantitative validation and broader object coverage are needed before the approach can be considered robust.

References

[1] Thomas E. Blanford, David P. Williams, Joshua D. Park, Benjamin T. Reinhardt, Kyle S. Dalton, Sean F. Johnson, and Daniel C. Brown. An in-air synthetic aperture sonar dataset of target scattering in environments of varying complexity. *Scientific Data*, 11:1196, 2024. doi: 10.1038/s41597-024-04050-0.

[2] Shirui Cao, Dong Li, Sunghoon Ivan Lee, and Jie Xiong. PowerPhone: Unleashing the acoustic sensing capability of smartphones. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2023. doi: 10.1145/3570361.3613270.

[3] Cheng Li, Jian Wang, Xiaobo Ding, and Ning Zhang. Acoustic imaging using the built-in sensors of a smartphone. *Symmetry*, 13(6):1065, 2021. doi: 10.3390/sym13061065.

[4] Xinyu Liang, Zikun Wei, Dong Li, Jie Xiong, and Jeremy Gummeson. SONDAR: Size and shape measurements using acoustic imaging. In *Proceedings of the Twenty-Fifth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc)*, pages 361–370, 2024. doi: 10.1145/3641512.3686359.

[5] Wenguang Mao, Mei Wang, and Lili Qiu. AIM: Acoustic imaging on a mobile. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2018. doi: 10.1145/3210240.3210325.

[6] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. FingerIO: Using active sonar for fine-grained finger tracking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016. doi: 10.1145/2858036.2858580.

[7] Sanjib Pradhan, Ghufran Baig, Wenguang Mao, Lili Qiu, Guobin Chen, and Bo Yang. Smartphone-based acoustic indoor space mapping. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):75, 2018. doi: 10.1145/3214278.

[8] Albert W. Reed, Jeonghyun Kim, Thomas Blanford, Aditya Pediredla, Daniel C. Brown, and Suren Jayasuriya. Neural volumetric reconstruction for coherent synthetic aperture sonar. *ACM Transactions on Graphics*, 42(4):113, 2023. doi: 10.1145/3592141.

[9] Piotr Serafin, Magdalena Okoń-Fąfara, Mateusz Szugajew, Czesław Leśnik, and Adam Kawalec. 3-D inverse synthetic aperture sonar imaging. In *2017 18th International Radar Symposium (IRS)*, 2017. doi: 10.23919/IRS.2017.8008209.

[10] Oliver Shih and Andrew Rowe. Can a phone hear the shape of a room? In *Proceedings of the 18th International Conference on Information Processing in Sensor Networks (IPSN)*, 2019. doi: 10.1145/3302506.3310407.

[11] Anran Wang and Shyamnath Gollakota. MilliSonic: Pushing the limits of acoustic motion tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019. doi: 10.1145/3290605.3300248.

[12] Sheng Wang, Xiuzhen Wang, Wei Jiang, Chao Miao, Qun Cao, Hao Wang, Ke Sun, Haoxiang Xue, and Lu Su. Towards smartphone-based 3D hand pose reconstruction using acoustic signals. *ACM Transactions on Sensor Networks*, 20(5):106, 2024. doi: 10.1145/3677122.

[13] Wei Wang, Alex X. Liu, and Ke Sun. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2016. doi: 10.1145/2973750.2973764.

[14] Bangyu Zhou, Mohammed Elbadry, Ruipeng Gao, and Fan Ye. BatMapper: Acoustic sensing based indoor floor plan construction using smartphones. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, pages 42–55, 2017. doi: 10.1145/3081333.3081363.

[15] Yongpan Zhuang, Yongjian Wang, Yuliang Yan, Xuhai Xu, and Yukang Shi. ReflecTrack: Enabling 3D acoustic position tracking using commodity dual-microphone smartphones. In *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2021. doi: 10.1145/3472749.3474805.