Poster: Provenance-Anchored Multi-Modal Authentication for Emergency Robot Overrides

Yuxin "Myles" Liu Nader Sehatbakhsh
Electrical and Computer Engineering, UCLA, Los Angeles, CA, USA
{liuyuxin@ucla.edu, nsehat@ee.ucla.edu}

***Abstract*—Autonomous mobile robots (AMRs) in public spaces need localized, tamper-resistant emergency overrides for first responders. Cloud access control fails under network partition, and unimodal credentials fall to relay, projection, and cross-modal association (Mafia-fraud) attacks. We bind an operator's cryptographic identity to their physical command vector by cross-attesting secure RF ranging (IEEE 802.15.4z UWB, Bluetooth 6.0 Channel Sounding), depth-SLAM human geometry, and directional acoustics, each signed at capture by a sensor root of trust. We then identify a residual attack that bearing-only acoustics cannot resolve: an adversary speaking from inside the direction-of-arrival (DoA) cone of an authenticated officer. This region is a beam, not a clinch, $0.94\,\mathrm{m}$ wide at $4\,\mathrm{m}$ for our measured 2.24° DoA RMSE, and unbounded in range, so we replace the existential binding predicate with an *exclusive-attribution* predicate that fails closed unless exactly one token-bound speaker can explain the ray. Our Agibot D1 EDU prototype decides in $\leq 3.05\,\mathrm{ms}$.**

I. Introduction

AMRs are deploying into municipal spaces, where crises demand that first responders halt a robot or clear an access route immediately. Centralized PKIs fail under network blackout or jamming, and local unimodal mechanisms are physically fragile: RF beacons admit distance-shortening and relay attacks [1], while voice and vision succumb to acoustic redirection and 2D spoofing [2]. Naive aggregation is worse, exposing a *cross-modal association vulnerability* (physical Mafia fraud) [3]: an adversary in front of the robot vocalizes a command that is validated against a genuine token held by an officer occluded nearby.

We bind officer to command by verifying geometric coherence across three spatial verifiers: RF distance bounding with angle-of-arrival (AoA), volumetric 3D human localization via depth SLAM, and acoustic beamforming (with front RGB proposing human hulls rather than serving as a credential), all rooted in hardware sensor provenance [4]. We then confront an attack this construction creates: binding is only as sharp as its coarsest modality, and acoustics gives *bearing without range*. We quantify the resulting ambiguity, show it far exceeds the close-quarters struggle case usually discussed, and give a predicate and a degraded mode that fail safely inside it.

II. Threat Model & Security Assumptions

Adversary Capabilities. An active, co-located adversary forces an unauthorized override or denies legitimate responder control via RF manipulation (sniffing, replay, distance-shortening, relay) and middleware tampering (forged frames or stale telemetry from unhardened ROS 2 nodes); *presentation and acoustic projection*, injecting commands from spoofed angles using planar posters, optical projections, or directional loudspeakers; *Mafia fraud*, vocalizing in line-of-sight while an authorized officer is occluded elsewhere; and *co-located speaker substitution* (§III-B), speaking from within the DoA ambiguity beam of an officer, such as laterally adjacent, grappling during an arrest, or anywhere along their bearing.

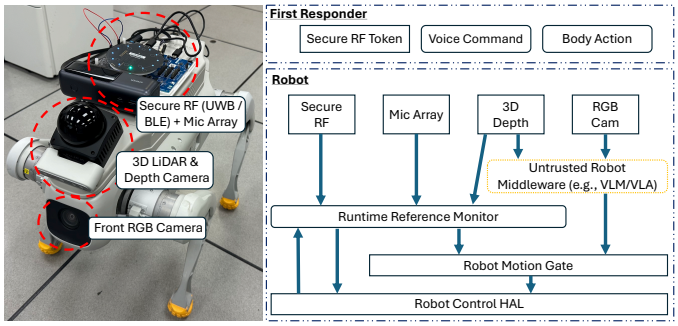


Fig. 1. (left) Prototype testbed; (right) system architecture abstraction.

Trusted Computing Base. Hardened wearable tokens (Secure Elements) performing authenticated distance bounding via IEEE 802.15.4z STS or Bluetooth 6.0 Channel Sounding; sensor roots of trust signing raw telemetry at capture time [4]; and an isolated Runtime Reference Monitor above the motor Hardware Abstraction Layer.

Out of Scope. Chassis or sensor destruction, cryptographic compromise, side-channel key extraction, token theft, and coercion of an authenticated officer; denial of officer mobility by physical body-blocking, whose integrity consequence is nonetheless handled in §III-B. Visual and acoustic streams are defended by geometric consistency, not adversarial ML robustness.

III. Provenance-Anchored Multi-Modal Binding

A. Cross-Modal Spatial-Identity Correlation

An isolated Runtime Reference Monitor decouples untrusted middleware (e.g., ROS 2 planners, VLM navigation) from the motor HAL, verifying three modalities before actuating a hardware-level Robot Motion Gate (Fig. 1). The officer's token distance-bounds via UWB STS or BLE 6.0 Channel Sounding, and the multi-antenna array yields a bounded range $r \pm \epsilon_r$ and AoA (θ, ϕ) spanning an RF cone $\mathcal{C}_{\mathrm{RF}}$. LiDAR and depth track human hulls $\mathcal{B}_{\mathrm{depth}}$ proposed by front RGB, where point density and surface-normal variance reject planar cutouts and

projections ($z \approx$ const). A microphone array beamforms the 3D DoA $\vec{v}_{\text{sound}}$ of the verbal command. To stop compromised nodes from injecting synthetic coordinates or replaying RF reports, roots of trust sign raw payloads with a counter k and capture timestamp: $\sigma_{\text{sensor}} = \text{Sign}_{sk}(H(\text{Payload}) \parallel k \parallel t_{\text{cap}})$ [4]; the monitor checks freshness and signatures before parsing topic data.

B. From Existential to Exclusive Attribution

A natural binding rule is *existential*: authorize if *there exists* a token-matched hull the ray could have come from, $\exists j : \vec{v}_{\text{sound}} \cap \mathbf{p}_{\text{head}}(\mathcal{B}_j) \neq \emptyset \wedge \mathcal{B}_j \cap \mathcal{C}_{\text{RF}} \neq \emptyset$. This defeats classical Mafia fraud but is unsound whenever more than one body can explain the same ray. Because a compact array measures bearing without range, the positions consistent with an observed DoA form a cone: unbounded along the bearing, of lateral width $w(d) \approx 2d\tan(3\sigma_{\text{DoA}})$ at range d. An adversary anywhere inside that beam inherits the officer's credential.

We therefore require *exclusive attribution*. Let $\mathcal{H}$ be the verified human hulls, $\mathcal{C}_{\text{snd}} = \{\mathbf{p} \mid \angle(\mathbf{p}, \vec{v}_{\text{sound}}) \leq 3\sigma_{\text{DoA}}\}$ the gated DoA cone, and $\mathcal{A} = \{\mathcal{B}_j \in \mathcal{H} \mid \mathbf{p}_{\text{head}}(\mathcal{B}_j) \in \mathcal{C}_{\text{snd}}\}$ the *ambiguity set*. The monitor authorizes iff

$$\mathcal{A} = \{\mathcal{B}^*\} \ \wedge \ \exists i : \ \mathcal{B}^* \cap \mathcal{C}_{\text{RF},i} \neq \emptyset, \tag{1}$$

i.e., exactly one body could have produced the utterance and that body holds a live credential. Eq. (1) strengthens the existential rule: it still rejects the occluded-officer relay, and additionally rejects the laterally adjacent and collinear adversary, including the arrest and grappling cases, with no new sensors. Multi-officer scenes use maximum-weight bipartite matching between cones and hulls; any unattributed body in the beam still contaminates the decision.

Availability under fail-closed. Suppressing ambiguous commands would otherwise hand the adversary a free denial of override: standing in the beam is enough to mute the officer. Voice is therefore a convenience channel: the token can also emit a signed, distance-bounded override with a trigger (e.g., a button press), so contamination suppresses vocal overrides only, never the officer's authority.

Restoring hands-free voice (ongoing). To re-admit vocal commands in contaminated clusters we are prototyping wearable-anchored attestation, signed through the same provenance chain: an accelerometer or contact mic responding to the wearer's own vocal-tract vibration but not to air-coupled speech at a set distance (e.g., 30 cm), defeating relay through the officer's own lapel; and an acoustic time-of-flight differential, cross-correlating a signed energy envelope from the wearable against the robot array to recover source range along the bearing. The two are complementary, as the latter is degenerate for an adversary exactly collinear behind the officer.

IV. Implementation & Preliminary Evaluation

Micro-benchmarks. We evaluate on an Agibot D1 EDU quadruped (ROS 2 Humble, 3D LiDAR, RGB-D camera, ReSpeaker XVF3800 4-mic array) with NXP SR250 UWB shields and Nordic nRF54L15 BLE 6.0 tags, under active depth-SLAM load. Table I shows UWB dominates BLE on ranging precision (6.3 vs. 79.7 cm RMSE). Spatial intersection is sub-microsecond (0.16–0.22 μs; P99 $< 0.4\,\mu$s) and acoustic ray-casting takes 1.65 μs; the budget is dominated by provenance verification and by querying DoA over USB (0.65 ms). Total decision overhead is ≤ 3.05 ms, fitting 20–50 Hz control loops without actuation jitter, and Eq. (1) adds only one cone test per tracked hull.

TABLE I
Ranging Profile and Authentication Overhead

Pipeline Stage / Metric	UWB (SR250)	BLE 6.0 (nRF54L15)
Ranging Accuracy (RMSE)	6.3 cm	79.7 cm
Provenance Overhead	2.4 ms	1.4 ms
Acoustic DoA Query	0.65 ms	0.65 ms
3D Spatial Intersect	0.16 μs	0.22 μs
Total Decision Overhead	**3.05 ms**	**2.05 ms**

Mafia fraud. The array achieved 2.24° DoA RMSE across arrivals $(0^\circ, 45^\circ, 90^\circ, 180^\circ)$. With an attacker shouting "Robot, stop!" at 1.5 m in front of the robot and an authorized officer 4.0 m away behind a partition, proximity-only validation accepted the command, while our monitor found the acoustic ray striking the attacker's hull, disjoint from $\mathcal{C}_{\text{RF}}$: the command was dropped in 3.05 ms, 100% rejection across trials.

Attribution ambiguity. Propagating the measured σ_{DoA} through the 3σ gate bounds the region in which acoustics cannot separate two speakers: lateral width $w(d) \approx 0.235\,d$, 0.35 m at 1.5 m, 0.94 m at 4.0 m, and unbounded along the bearing. The exposed volume is a beam, not a clinch, so an entanglement-triggered policy leaves most of it uncovered, whereas Eq. (1) rejects by construction anywhere inside it.

V. Ongoing Work & Poster Artifacts

Beyond wearable attestation (§III-B), we are instrumenting the collinear case (matched azimuth, $\Delta d = 1$–4 m) to measure rejection under both predicates, and handling sensor degradation in smoke or dust, where missed hull detections become a soundness risk for Eq. (1) and we fail closed on low depth confidence. At ACSAC we will present the testbed hardware and logged sensor traces, with split-screen video contrasting authenticated overrides against rejections of both Mafia-fraud relay and co-located speaker substitution; if ready, an interactive visualizer of the intersection between point clouds, RF cones, and the gated ambiguity beam.

References

[1] P. Staat, K. Jansen, C. Zenger, H. Elders-Boll, and C. Paar, "Analog physical-layer relay attacks with application to bluetooth and phase-based ranging," in *Proceedings of the 15th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 2022, pp. 60–72.

[2] A. Solanki, W. Al Amiri, M. Mahmoud, B. Swieder, S. R. Hasan, and T. N. Guo, "Survey of navigational perception sensors' security in autonomous vehicles," *IEEE Access*, vol. 13, pp. 104 937–104 965, 2025.

[3] J. Zhang, A. Yang, Q. Hu, and G. P. Hancke, "Security implications of implementing multistate distance-bounding protocols," in *Proceedings of the ACM Workshop on Cyber-Physical Systems Security & Privacy*, 2019, pp. 99–108.

[4] Y. Liu, Z. Yao, M. Chen, A. Amiri Sani, S. Agarwal, and G. Tsudik, "Provcam: A camera module with self-contained tcb for producing verifiable videos," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 2024, pp. 588–602.