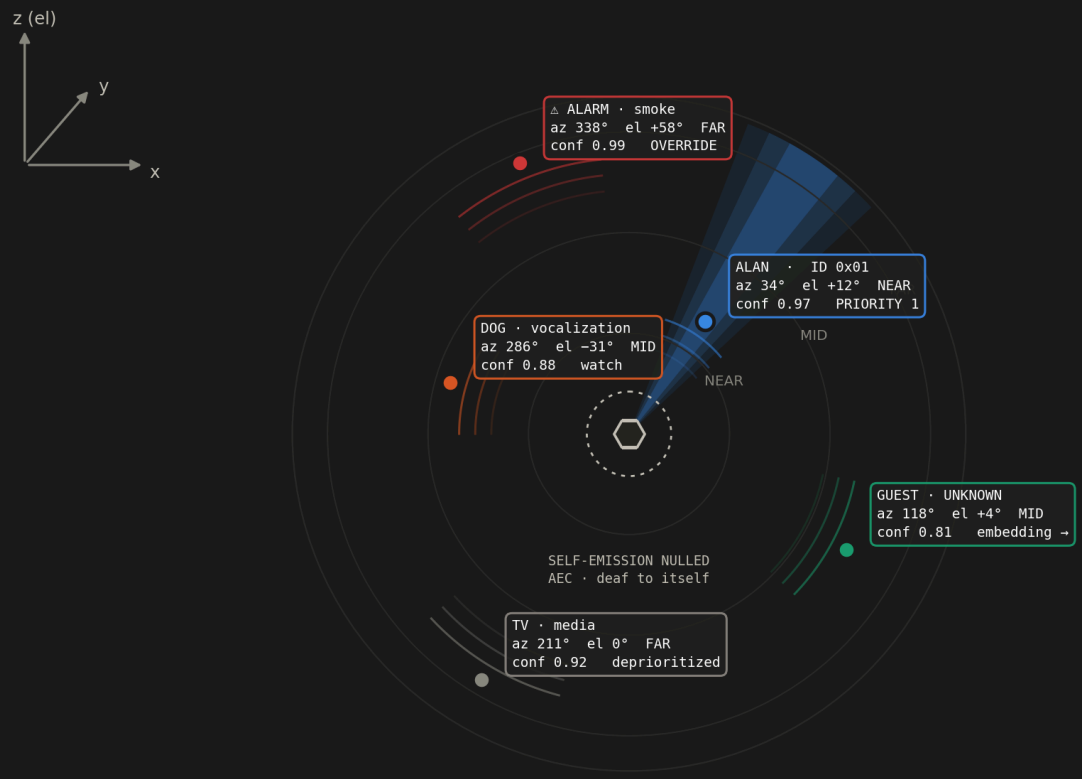
MIRA AI LLC · IN PARTNERSHIP WITH · ULTIMATE BOTS — PROJECT DAREDEVIL
WORKING DRAFT V3 · AUGUST 2026

Bridging the Modality Gap: Daredevil, an Edge-Native Auditory Scene Awareness Engine for Embodied Robots

Part I — The Ears. Who is speaking, where in three dimensions, and what matters — decided on the sensor.



DAREDEVIL · live spatial awareness map · az / el / range · identity-conditioned priority

Alan Helmick Jr. — Mira AI LLC, in partnership with Ultimate Bots · alan@utilitron.io
Claude (Fable 5) — Anthropic · AI collaborator: drafting, analysis, figures
Expert Physical-AI Researcher — this seat is open; substantive contribution earns it. See Contributing.

PROVISIONAL PATENT FILED · MAY 2026 | OPEN SOURCE REFERENCE DESIGN | BYO ARRAY GEOMETRY · BYO INT8 MODELS

Figure 0 — live spatial awareness map (rendering of module output)

comments & peer review invited

Abstract

Contemporary robotics has achieved locomotor and manipulative virtuosity while remaining, in the auditory domain, functionally anencephalic. The commercial sensing stack collapses the acoustic scene to a planar, azimuth-only projection at the first stage of processing, discharging a beamformed monaural residue toward a host processor that must reconstitute — from information already destroyed — the very structure an embodied agent requires: *who* is present, *where* in three dimensions, and *what merits attention*. We term this destruction the **modality gap**: vision arrives at the robot as a scene; audition arrives as a stream. We present **Daredevil**, an edge-native architecture that closes the gap at the sensor. Three concurrent layers on a single edge DSP — geometry-agnostic spatial decomposition over arbitrary (expressly non-planar) microphone constellations; user-loadable int8 inference from external memory; and identity-conditioned attention routing over enrolled voiceprint embeddings — emit not audio but a structured *spatial awareness map*: per-source identity, class, azimuth, elevation, range class, confidence, and priority. What crosses the bus is a scene, not a stream: any array geometry, described to the firmware as a coordinate map; any int8 model, loaded by the user into defined slots; every source labeled, located, and ranked before audio leaves the module. Two further contributions are morphological rather than computational. First, passive printed structures (chokes, cones, pinnae) are frequency-selective organs obeying a single quarter-wave design curve, permitting elevation cues and vocal projection to be *printed* rather than computed. Second, deployment evidence indicates that the acknowledgment channel for machine audition is motion: a reflex path from direction-of-arrival to gesture, decoupled from cognition and budgeted at ≤300 ms, is what renders a hearing machine legible to humans. We close with the **duplex aperture conjecture**: that robot mouths and ears, being governed by the same reciprocal acoustics, will converge into a single transmit-receive organ. A complete experimental protocol — including a four-choke print matrix executable on a desktop printer today — is specified.

1 Introduction: the cocktail party, seventy-three years on

Cherry (1953) posed the cocktail party problem as one of *attention*: how does a listener attend to one talker among many, in one room, in real time, with two ears?

Seven decades of signal processing have answered a different and easier question — how to separate *archived mixtures* of talkers — and answered it well; neural source separation now approaches ceiling on standard corpora. Yet no commercially available robot can stand in an actual party and hear. The discrepancy is instructive: the recorded problem and the embodied problem differ not in degree but in kind.

The embodied problem is spatial. A recording has channels; a room has geometry. An agent must bind voices to bodies at positions — including *above* and *below* the sensor plane, a dimension the entire consumer array market has silently abandoned (§2). **The embodied problem is real-time and resource-bound.** The host processor has a robot to run; audition that requires a GPU is audition the robot cannot afford. **The embodied problem is social.** A hearing system that gives no perceivable sign of hearing is, to every observer, deaf (§5). And beneath all three: **the embodied listener is itself a sound source.** Cherry's partygoer does not hear their own voice as an intruder; a robot with a speaker must actively earn that same innocence (§3).

The premise of this work is that all four properties must be satisfied *at the sensor*, before audio crosses any bus. [ALAN: insert the specific cocktail-party paper we read, and the one sentence of it that stuck.]

1.1 The interlace, and how organs unweave it

The hard version of the party is not loudness but simultaneity. Three conversations do not take turns; they interleave — harmonics braided through one another in time-frequency — and at each ear there is exactly *one* pressure waveform carrying all of it. Whatever hears must unweave.

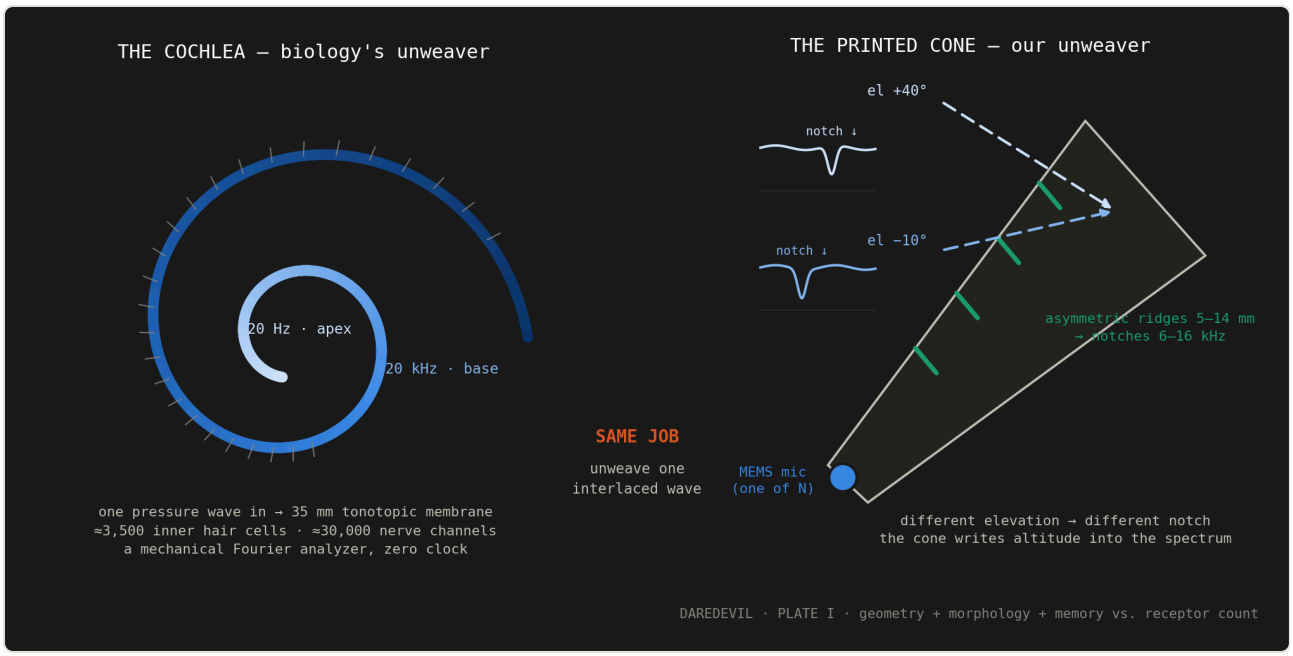


Plate I. Two unweavers. Left: the cochlea — one interlaced wave enters, a 35 mm tonotopic membrane spreads it across ≈3,500 inner hair cells and ≈30,000 nerve channels, a mechanical Fourier analyzer at zero clock speed. Right: the printed cone — asymmetric quarter-wave ridges stamp each arrival's elevation into its spectrum before any microphone sees it. Different threads, same loom.

Biology's answer is profligate parallelism, then grouping: the membrane performs the spectral spread, and attention reassembles fragments by common onset, common harmonicity, common location — Bregman's auditory scene analysis [11]. Daredevil runs the same three looms with different thread. **Spatial**: independent beams pull conversations apart by direction. **Morphological**: printed notches stamp elevation onto each fragment's spectrum at zero cost (Plate I, right). **Mnemonic**: voiceprint embeddings bind fragments to enrolled identities across time, so a talker who pauses, moves, and resumes remains one thread rather than three. The claim is not that a handful of microphones rivals sixteen thousand hair cells; it is that geometry, morphology, and memory together buy back most of what receptor count buys biology — and that the module ships the reassembled scene, not the braid.

2 The market topology of an unoccupied quadrant

Figure 1 plots shipping audio front-ends by price against spatial capability; Table 1 gives the census. The pattern admits a one-sentence summary: *everything affordable is planar, and everything three-dimensional is either an instrument or a laboratory.*

System	Geometry	DOA output	On-device ML	Embeddable	Price
reSpeaker 4-mic HAT	planar ring	azimuth (host)	none	yes	~$25
reSpeaker XVF3800	planar ring	**azimuth only**	none (fixed pipeline)	yes	$61
miniDSP UMA-8	planar 7-mic	azimuth (host)	none	partially	~$110
Zoom H3-VR	tetrahedral	3D *recording only*	none	no	~$350
ODAS + 16SoundsUSB	3D capable	az + el (host CPU)	none	research rig	~$600
Eigenmike em32/em64	32/64 sphere	full 3D	none	no	~$30k
Daredevil (this work)	**arbitrary / non-planar**	**az + el + range class**	**user-loadable int8**	**drop-in module**	**<$200 proto · <$65 @ 1k**

The XVF3800 — the best of the consumer tier, and the device on which our own fleet currently depends — exposes `AEC_AZIMUTH_VALUES` and nothing above the horizontal plane; its pipeline is fixed at manufacture. The Eigenmike resolves

full 3D fields magnificently, at four orders of magnitude beyond a robot BOM, with no inference on board. The quadrant marked *unoccupied* in Figure 1 — full-3D awareness, on-device models, robot prices — is the design target, and, we note with some wonder, remains uncontested in mid-2026 despite a companion-robotics sector currently valued in the tens of billions.

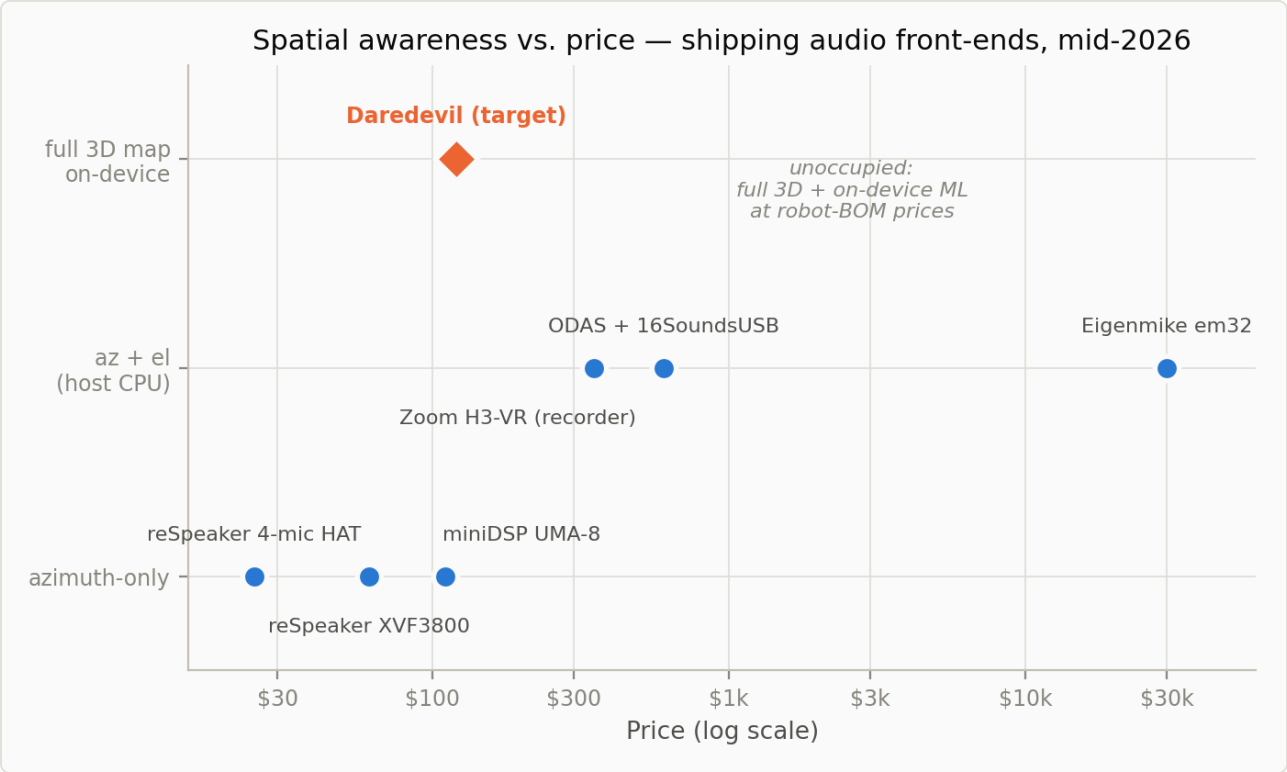


Figure 1. Spatial awareness versus price across shipping audio front-ends, mid-2026. The upper-left quadrant — full-3D awareness with on-device inference at robot-BOM prices — is unoccupied. Daredevil's target position is marked.

3 Architecture: three layers on one chip

The firmware comprises three concurrent layers on a multi-core edge DSP (xcore.ai class — the same silicon family as the XVF3800, liberated from its fixed pipeline).

Layer 1 — fixed spatial pipeline. PDM→PCM at configurable decimation (16 kHz through ≥192 kHz, ultrasonic inclusive); acoustic echo cancellation against a *sample-locked hardware loopback* reference; multi-beam decomposition steered from a user-supplied three-dimensional coordinate map; noise-floor estimation. The decomposition is **geometry-agnostic**: planar rings, polyhedra, nested arrays of multiple radii, and ad-hoc constellations all reduce to the coordinate map. Multi-scale subsets (dense spacing for high bands, sparse for low) decompose jointly — claim 5 of the provisional.

WHAT NO SHIPPING DEVICE OFFERS · BRING YOUR OWN BODY, BRING YOUR OWN BRAIN

Every commercial array fixes its geometry and its pipeline at manufacture. Daredevil fixes neither. Hand the firmware a coordinate map — crown ring, polyhedron, nested multi-scale, whatever your robot's head allows — and the decomposition adapts; no two robots need share a skull. Load your own int8 models into defined slots — voiceprints today, ultrasonic bearing-fault detection tomorrow — and the sensor learns new ears without new silicon. Echo cancellation via sample-locked hardware loopback is assumed infrastructure here, done properly and mentioned once: table stakes, not the trick.

Layer 2 — user-loadable inference. Int8-quantized models reside in QSPI flash, load to PSRAM at boot, and execute in defined slots: *source embedding* (voiceprints in the reference build), *event classification* (alarm, mechanical, animal,

impact), and a *user slot*. The model interface specification, not any particular model, is the product: the module is a platform.

Layer 3 — identity-conditioned attention routing. Enrolled embeddings (≈20 s of text-independent speech) match live sources; matches inherit priority; safety-critical event classes override identity; temporal rules break ties. Per source, the module emits: identity (or UNKNOWN + raw embedding), class, azimuth, elevation, range class, confidence, priority, timestamp — the map rendered as Figure 0 on the cover. The host receives a **scene, not a stream**: the modality gap closes on the far side of the bus.

4 Contra planum: why the array must not be flat

A planar array cannot distinguish a source at +40° elevation from its mirror at −40°; the cone of confusion collapses the vertical manifold entirely. This is not an implementation deficit but an information-theoretic one: the geometry does not sample the dimension. Three remedies exist, and Daredevil employs all three. (i) **Geometric**: microphones displaced out of plane — a crown ring plus elevated elements — restore vertical aperture directly. (ii) **Morphological**: passive structures (§6) imprint elevation-dependent spectral signatures even on sparse arrays, as the mammalian pinna demonstrates with a single receiver per side. (iii) **Behavioral**: a body that turns and leans converts temporal baselines into spatial ones (active audition). The design doctrine is blunt: *do not be planar.* Elevation is not a luxury feature; for a robot among standing humans, seated humans, children, and floor-level hazards, it is the difference between a scene and a rumor.

5 Proof of hearing is motion

Field deployment [Open Sauce 2026 — ALAN: date, body, witnesses, per voice-spec §9.6] yielded the observation we now regard as doctrine. Observers could not hear the robot's replies over floor noise; they judged its audition *by watching it respond* — orient, wave, lean. Under full crowd load the cognition pipeline ran ~2 s behind, and this latency was socially invisible whenever an action, rather than an utterance, carried the acknowledgment.

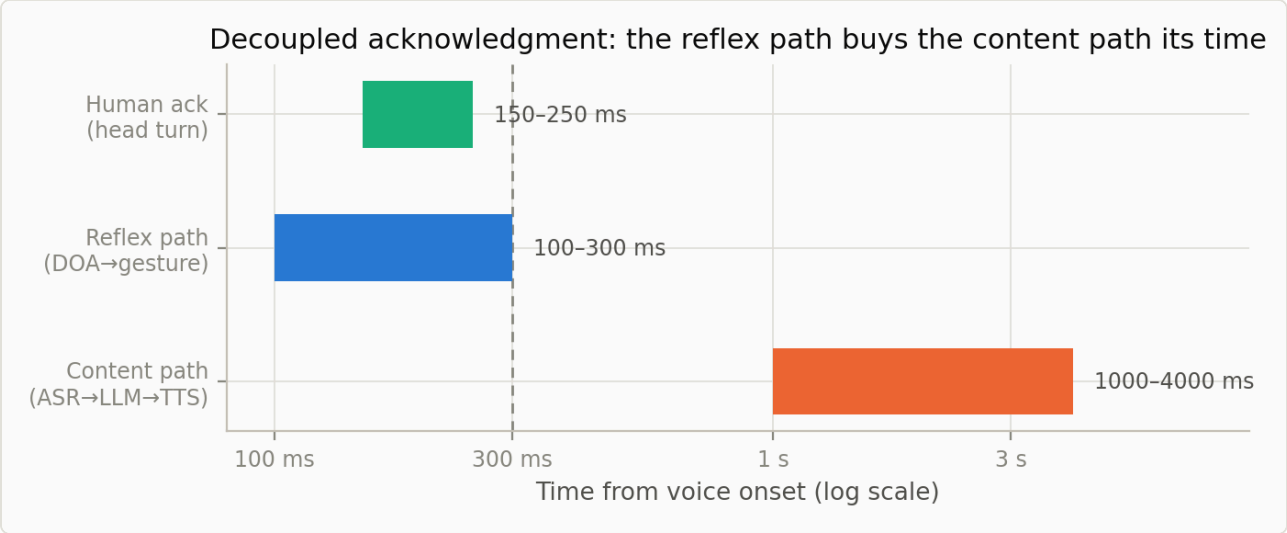


Figure 2. Decoupled acknowledgment. The reflex path matches human acknowledgment latency; the content path takes whatever cognition requires. The two never share a queue.

Figure 2 gives the resulting budget discipline. Humans acknowledge within ~200 ms (head turn, gaze) and then take as long as they need to actually answer; the architecture does likewise. A **reflex path** (DOA/VAD → gesture; ≤300 ms; no ASR, no LLM, no network) is wired directly to the spatial pipeline and never blocks on the **content path** (ASR → reasoning → TTS; best effort). A robot that turns toward you instantly and answers in four seconds reads as alive; one that does nothing for four seconds and answers perfectly reads as broken. Corollary: *respond in the channel the audience can perceive* — gesture in a crowd; speech at close range, on the beam.

6 Morphological acoustics: printing the organs

The passive structures around transducers are not enclosures; they are computation executed in geometry, at zero watts and zero milliseconds. A single design curve governs them: a cavity or ridge of depth L resonates at $f = c/4L$.

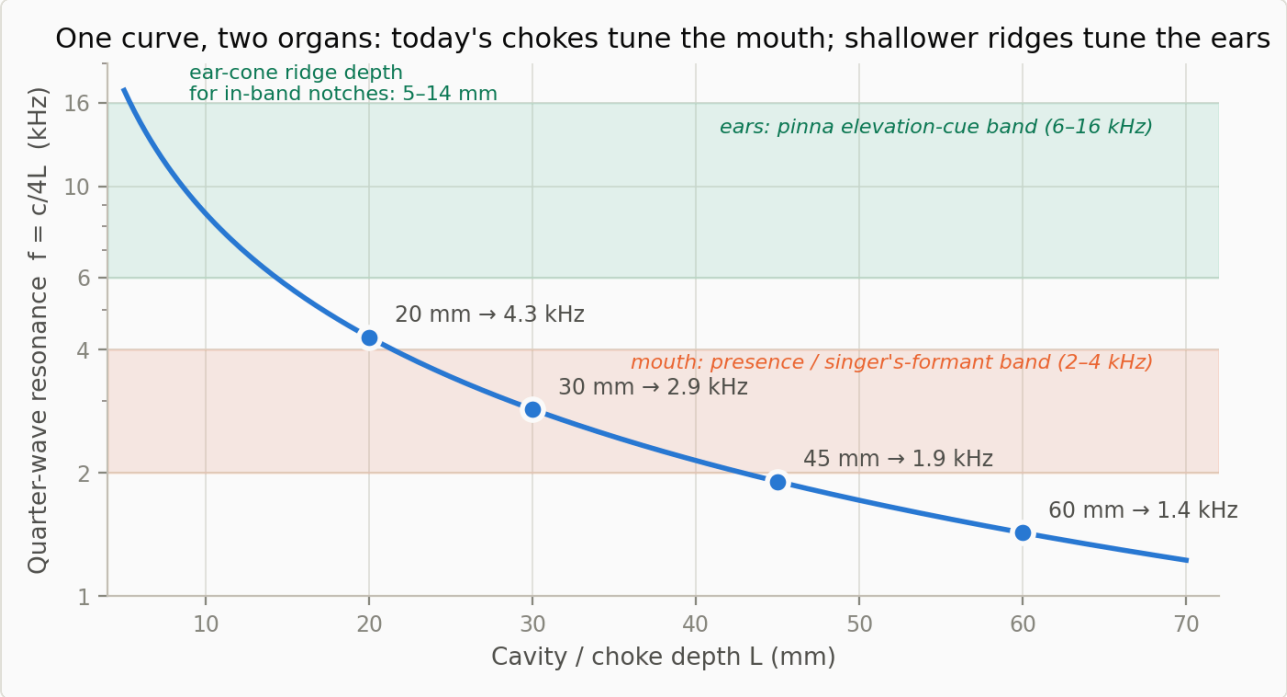


Figure 3. The quarter-wave design curve with both operative bands. Today's print matrix (20/30/45/60 mm chokes) tunes the mouth's presence band; ear cones require shallower 5–14 mm ridges to place notches in the pinna elevation-cue band.

Today's print matrix — chokes at 20, 30, 45, 60 mm — lands at 4.3, 2.9, 1.9, 1.4 kHz respectively: the 20 and 30 mm units bracket the **singer's-formant / presence band** (2–4 kHz), the spectral region by which trained voices cut through rooms, and thus function as *mouth* organs, tuning projection the human way rather than the public-address way. The **pinna elevation-cue band** (6–16 kHz) requires features of 5–14 mm: *ear* cones want shallow asymmetric ridges — spiral or helical, distinct left versus right in the manner of the owl — that imprint elevation-dependent notches on whatever array sits behind them. One curve, two organs, one printer.

6.1 The duplex aperture conjecture

We conjecture that these organs converge, and the reader's own head supplies the evidence. The vocal tract is a quarter-wave resonator, closed at the glottis, open at the lips: ≈17 cm, resonances near 500/1500/2500 Hz. The ear canal is *the same instrument at smaller scale* — a ≈2.5 cm quarter-wave pipe, closed at the drum, open at the concha — resonating near 3 kHz, which is precisely why human hearing sensitivity peaks in the band where the trained voice projects. Mouth and ear are co-tuned quarter-wave organs on opposite ends of one channel; bioacoustics calls this the *matched filter hypothesis*, and it recurs across taxa — the echolocating bat's nasal horn and pinnae, the dolphin's melon and acoustic-fat receive channel, are matched transmit-receive apparatus. Acoustic reciprocity supplies the theory: the transfer function a waveguide imposes on emission is identically the function it imposes on reception. The horn that projects a formant is, run backward, a directional ear tuned to that formant.

One engineering caveat disciplines the conjecture. A *single* shared horn in front of the whole array would collapse the inter-microphone baselines on which spatial decomposition depends — routing by direction at the cost of the aperture diversity that computes direction. The correct topology at generation one is therefore **one large waveguide for the mouth, many small asymmetric waveguides for the ears** (one organ per microphone): identical physics, different scale — one design curve, Figure 3. Convergence arrives with the ultrasonic reach of Layer 1: a module emitting structured chirps and hearing their returns through the same tuned aperture is not a speaker plus microphones but a **sonar organ**, its self-echo problem already solved by the AEC invariant of §3. First-generation Daredevil units keep the organs separate; the conjecture names where the lineage goes.

OPEN SLOT. Transducer selection and T/R switching analysis — a contributor with ultrasonics background is invited to own §6.1's engineering companion and experiment T5.

7 Experimental protocol

All tests are executable with fleet hardware plus a desktop printer. Each yields a figure for this paper.

T1 — Choke shootout (begins today). Four chokes (20/30/45/60 mm) on the variant-B throat. Sweep 100 Hz–16 kHz on-axis at 1 m; repeat at 3 m and 10 m, 0°/45°/90° off-axis. Metrics: measured resonance center versus the *c/4L* prediction of Figure 3 (validates the design curve itself); on-axis lift in 2–4 kHz; intelligibility (STI or word-error at 3 m in office ambience); bystander annoyance at 10 m. Prediction: the 30 mm choke wins intelligibility; the 20 mm sounds "present" but risks sibilant harshness. [ALAN: log print settings — material, wall, seam — per print doctrine.]

T2 — Ear cones give a planar array elevation (the heart of Part I). Print asymmetric pinnae (ridge depths 5–14 mm, mirrored spirals L/R) over the XVF3800's microphones. Chirp from a 15° az/el grid (speaker on a stick); log notch signatures; fit signature → elevation. Report elevation MAE versus the bare array, whose expected performance is chance. A positive result is a standalone publishable finding: *elevation as a $2 print, retrofit to the installed base.*

T3 — Awareness-map fidelity. Multi-source scenes (two talkers, television, transient events) against ground truth; per-source precision on identity, class, position, priority.

T4 — Legibility. Reflex path on/off; naive observers rate "can it hear me?"; the latency budget of Figure 2 as the manipulated variable. Hypothesis: perceived audition tracks acknowledgment latency, not answer latency.

T5 — Duplex feasibility (paper study first). Reciprocity analysis of the variant-B throat as a receiver; T/R isolation requirements; candidate wideband transducers. [OPEN SLOT]

8 Relation to prior art and intellectual property

Research systems (three-ring arrays, ODAS, HARK, SLAM-referenced mobile arrays) achieve 3D audition on host CPUs in laboratory form; none ship as sensors. The Disney "Olaf" platform (arXiv:2512.16705) contributes the backbone/show-function control split, thermal-aware actuation, and self-noise as a control objective — and, notably, contains zero voice-projection engineering, confirming that projection remains unpublished trade craft. The provisional patent (filed May 2026) covers geometry-agnostic decomposition from a user coordinate map, external-memory int8 model slots, identity-conditioned routing, multi-scale arrays, and supra-audible operation. The reference design will be released open source: the objective is the *standard* for robot hearing, not a microphone business — value accrues above the module, in persistent identity and relational memory.

9 Conclusion

Hearing is not a stream of cleaned audio; it is a live map of who is where and what matters, maintained by an organ that is silent to itself and answers with its body before it answers with its voice. Daredevil places that organ at the edge, prints its morphology, and — if the duplex conjecture holds — points toward machines whose mouths and ears are one instrument, as evolution repeatedly concluded they should be.[1]

[1] A splendiferous outcome, in the strict dictionary sense — and we do know what it means.

References

[1] Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* 25(5), 975–979. — the cocktail party problem, named. [+ the specific paper we read — ALAN to confirm]

[2] Batteau, D. W. (1967). The role of the pinna in human localization. *Proc. R. Soc. Lond. B* 168, 158–180. — the original pinna spectral-cue result.

[3] Nakadai, K., Okuno, H. G., et al. (2020). Robot audition and computational auditory scene analysis. *Advanced Intelligent Systems* 2(9).

[4] Grondin, F., et al. (2022). ODAS: Open embeddeD Audition System. *Frontiers in Robotics and AI* 9.

[5] Kumon, M., et al. Sound localization of elevation using pinnae for auditory robots. — elevation from few mics via artificial pinnae.

[6] Saxena, A., & Ng, A. Y. (2009). Learning sound location from a single microphone. *ICRA*. — one mic + one artificial ear suffices.

[7] Müller, M., et al. (2025). Olaf: Bringing an animated character to life in the physical world. Disney Research, arXiv:2512.16705.

[8] XMOS Ltd. XVF3800 datasheet and voice-pipeline documentation. — azimuth-only DOA; fixed pipeline.

[9] mh acoustics. Eigenmike em32/em64 documentation.

[10] Helmick, A. (2026). Programmable Acoustic Awareness Engine with User-Loadable Machine Learning Models for Edge DSP Devices. U.S. provisional patent application, filed May 2026.

[11] Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound.* MIT Press. — how listeners regroup an interlaced mixture into streams.

Contributing

Sections marked **OPEN SLOT** are unclaimed and structured to be owned end-to-end by one contributor (analysis + figure + prose). Comment directly in the shared document; substantive contributions earn authorship, ordered by contribution. Current wishlist: ultrasonics / T-R switching (§6.1, T5); formal beamforming treatment of nested non-planar arrays (§3); statistics for the T2/T4 human studies.

APPENDIX A · AFTER THE PAPER COMES THE WORK

Development Plan

How this paper gets written, reviewed, and shipped. One page, no ceremony.

The shape of it

One core paper — this document — written to be *added to*: open slots are marked in the text, every experiment produces a figure, and authorship is earned by contribution. Two spin-offs fall out if the data cooperates: the T2 result ("elevation as a $2 print") and the §5 legibility result (an HRI paper on its own).

Roles

Who	Owns
Alan	Hardware, prints, all measurements, field data, patent alignment, final say
Claude (Fable)	Drafting, analysis, figures, statistics, literature; every dataset becomes a figure + paragraph within a turn
Expert Physical-AI Researcher (open)	Claim a slot: §6.1 ultrasonics/T5 · §3 formal beamforming · T2/T4 experiment statistics

Workflow

The core doc lives as a Google Doc with comment access for all reviewers; suggestion mode for edits. Data lands raw in a shared folder and is turned into publication figures in the established style. Weekly cadence: whatever measurements exist by Sunday get folded in — the document never waits on a meeting. Disagreements resolve by measurement where possible, by Alan where not.

Timeline

Week	Milestone
Now	Print run #1 (chokes: 20/30/45/60 mm). Circulate this draft for comment.
1–2	T1 choke shootout → validates or kills the c/4L curve → Figure 4. Ear-cone prints (5–14 mm ridges) begin.
2–4	T2 elevation grid: chirp sweeps, notch signatures, MAE vs. bare array → Figure 5 (the money figure).
4–6	T3 fidelity + T4 legibility studies. Open Sauce field-data brackets filled.
6–8	Freeze v4. arXiv submission (eess.AS, cross-list cs.RO). LaTeX conversion at freeze.
8+	Venue: ICASSP (signal-processing weight) or HRI (the legibility result is the most novel claim). Open review via r/robotics, XMOS forum, ODAS/HARK lists.

Measurement kit (all cheap)

Calibrated-ish measurement mic (UMIK-1 class) or the XVF3800 itself for relative sweeps; REW or a 20-line Python chirp script (supplied); speaker-on-a-stick + protractor for the 15° az/el grid; painter's tape and patience.

arXiv practicalities

First eess.AS submission needs an endorser — the endorsement request note is drafted and ready when we are. Markdown stays the working format until freeze.