INSTITUTE FOR PHYSICAL AI @ BMI
The Charlot Lab

Technical Report TR–2026–22
Survey / Review · Preprint v1
21 July 2026

TOUCHING SOUND

Touching Sound: Contact Audio and the Complementarity of Physical Perception

The sense that reaches through occlusion and into the sealed interior — and why sensor fusion is coverage, not redundancy.

David Jean Charlot, PhD

Dean of Physical AI · The Charlot Lab, Institute for Physical AI @ BMI

Correspondence: contact@physicalai-bmi.org · physicalai-bmi.org
Interactive companions: physicalai-bmi.org/research/charlot-lab/topic/acoustic

Abstract. Vision resolves an object's shape and a fingertip resolves its texture, but neither can tell a solid billet from a hollow shell, water from air behind a closed wall, or a seated bearing from a cracked one, and both go dark under occlusion or in the dark. Contact audio can. When a hand taps, scratches, or shakes an object, the vibrations it radiates carry precisely what light and touch cannot reach: material, internal structure, and fill level, sensed through the wall. This report reviews the emerging use of contact microphones in robotic perception, from in-hand acoustic object recognition to in-gripper audio for contact-rich manipulation, and situates it within the broader program of audio–visual–tactile fusion. Its central argument is that the dominant justification for fusing senses — averaging away noise on a shared quantity — understates the case: the sharper reason is *coverage*. Some properties of the physical world have exactly one sensory witness, and for what is sealed, hollow, or hidden, that witness is sound. We formalize this with a per-attribute Bayesian model in which each sense contributes a likelihood from its own confusion structure, and show that removing the acoustic channel collapses the posterior over fill level to chance while leaving the other attributes intact. Section 2 states the review method. Sections 3–5 develop the physics of contact sound, the sensing hardware, and learning from it. Section 6 develops the complementarity argument. Section 7 describes three transparent, on-device companion instruments that make the argument legible. The report reports no new experimental measurements; the companions are didactic models, numerically verified as described, not empirical results.

1. Introduction

A remote sensor reports what arrives from a distance: photons for a camera, range for a depth sensor, a pressure field at a diaphragm for a room microphone. A contact sensor reports the mechanical state of an interface: how deep an object has pressed into a finger, whether the contact is holding or sliding, the micron texture at the point of touch. Between these two regimes sits a class of object properties that neither can reach. The fill level of a closed bottle, the hollowness of a casting, the crack inside a weld, the seating of a valve: these are defined *inside* the object, behind a wall that light cannot penetrate and a fingertip cannot feel past. A camera looking at an opaque, sealed bottle sees exactly one thing — a bottle — whether it is empty, half full, or full.

Contact audio reaches these properties because it is not a look at the object from outside but a readout of the object's own mechanics. Strike a body and it vibrates in its natural modes; the frequencies of those modes are set by its stiffness and mass distribution, and the rate at which they decay is set by its internal damping. A struck metal cup rings bright and long; a struck block of wood answers with a short, dark thud; a bottle's pitch shifts as it fills. To tap an object is to excite this signature and to hear, in it, the material and the interior. This is not a metaphor: the informative signal is literally composed of the property one wants to know.

Human perception already fuses these channels — shape by sight, material by sound, surface by touch — and a growing body of robotics work now instruments the same combination, placing microphones on fingertips and in grippers so that a manipulator can hear what it handles[1, 2]. This report reviews that work and argues for the reason it matters most, which the noise-averaging account of fusion tends to hide: for hidden properties, sound is not a better view of the visible — it is the only view of the sealed.

2. Review method and scope

This is a narrative review. It covers three threads: contact-microphone hardware and in-hand acoustic object recognition; the use of in-gripper audio to learn contact-rich manipulation; and audio–visual–tactile fusion for scene understanding. It draws the physics of contact sound from classical acoustics[6, 7] and the fusion argument from probabilistic robotics[9]. The report introduces no new datasets, hardware, or experimental measurements. Where it states quantities — resonant frequencies, classification accuracies, posterior collapse — these are outputs of the small, transparent models described in Section 7, each verified numerically before it was built into an interactive companion, and each labeled as a didactic construction rather than an empirical finding.

3. What a tap radiates

Consider an object excited by a brief contact — a tap, a scratch, a shake. Three quantities of the radiated sound carry the properties of interest.

Material, from the modal spectrum. A struck body vibrates in a set of natural modes whose frequencies depend on the elastic moduli and density of the material and on the geometry. Two useful, geometry-robust summaries are the distribution of energy across frequency and its brightness, the spectral centroid

$$c = \frac{\sum_k f_k \, |X_k|}{\sum_k |X_k|},$$

where $|X_k|$ is the magnitude of the sound at frequency f_k. Metals concentrate energy at high frequencies and ring with a high centroid; wood and plastic damp the highs and answer darker. A second summary is the ring-out time, the interval over which the energy decays, which reports internal damping and therefore, indirectly, whether a body is solid or hollow.

Fill and hollowness, from resonance shift. The most striking demonstration that sound reaches the interior is a filling vessel, whose pitch changes with a level no camera can see through an opaque wall — and, tellingly, changes in opposite directions depending on how it is excited. Blow across the open mouth and the enclosed air acts as the spring of a Helmholtz resonator, so the resonant frequency is

$$f_{\mathrm{H}} = \frac{c_{\mathrm{air}}}{2\pi} \sqrt{\frac{A}{V\, L'}},$$

with A the neck cross-section, L' its effective length including the end correction, and V the enclosed air volume[7]. As liquid fills the vessel, V shrinks, the air spring stiffens, and the pitch *rises*. Strike the wall instead and the vessel rings as a shell loaded by the liquid moving with it; treating the wall as a spring–mass system whose stiffness is fixed and whose participating mass grows with the wetted height gives, to first order,

$$f_{\mathrm{S}}(\phi) = \frac{f_0}{\sqrt{1+\alpha\,\phi}},$$

where ϕ is the fill fraction and α a dimensionless mass-loading coefficient. Here more liquid adds mass to the same stiffness, so the pitch *falls* — the familiar effect of tapping a row of glasses filled to different levels[8]. Figure 1 plots both. The point is not the specific numbers but their monotonicity: in either mode, an unseeable interior state maps one-to-one onto an audible pitch.

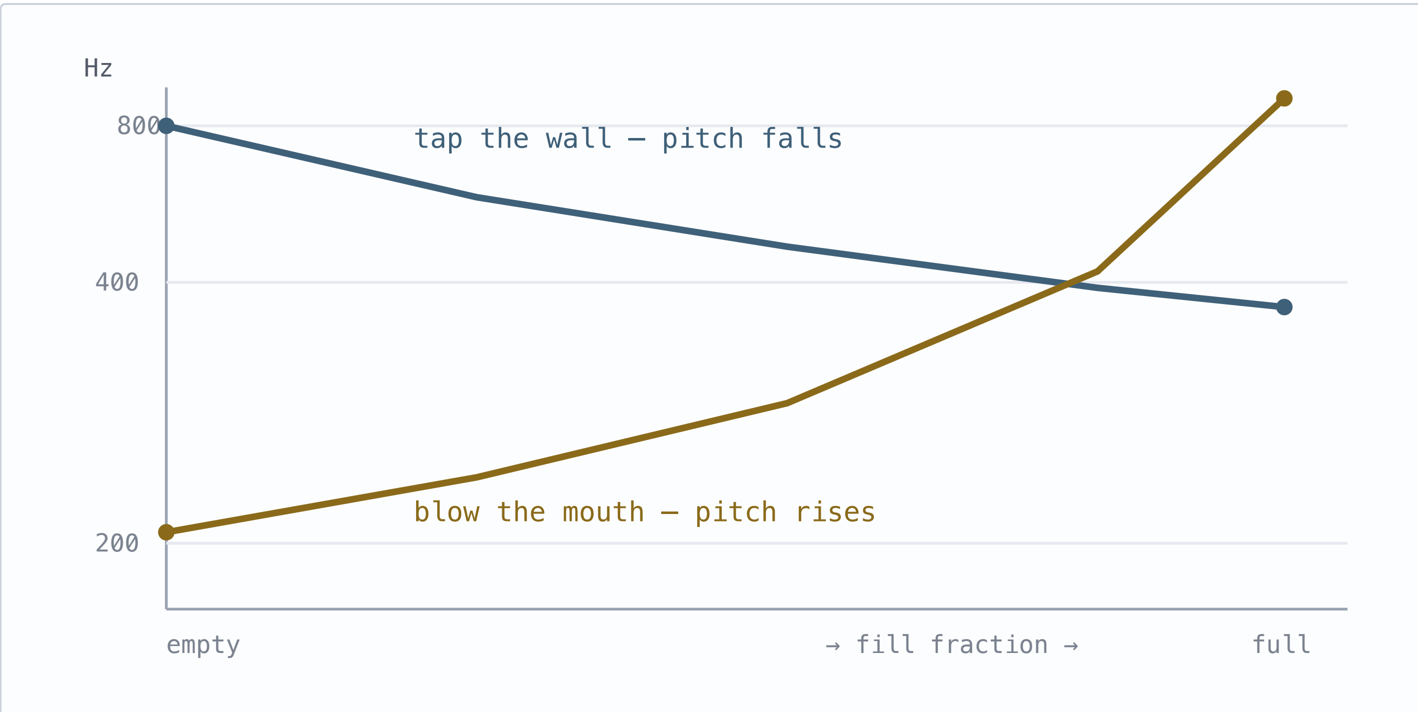

Figure 1. Resonant pitch versus fill fraction for a ~500 mL vessel, on a logarithmic frequency axis. Blowing across the mouth excites a Helmholtz air resonance that *rises* as the air volume shrinks; tapping the wall excites a liquid-loaded shell mode that *falls* as mass is added. The two mechanisms move in opposite directions, yet each renders the hidden fill level as an audible pitch. Curves computed from the closed forms in Section 3 ($\alpha = 1.78$, chosen so a full vessel rings at ~0.6× the empty pitch).

4. Sensing hardware: microphones at the point of contact

The sensor that reads contact sound is not the room microphone but the contact microphone: a piezoelectric or MEMS element coupled to the skin or structure, which responds to structure-borne vibration rather than to airborne pressure. Placed on a fingertip or inside a gripper, it hears the object being handled and rejects most of the ambient field, because the informative vibration arrives through the mechanical contact, not through the air. SonicSense equips a multi-fingered hand with such microphones and demonstrates that, purely from the vibrations of tapping, scratching, and shaking, a robot can identify materials and recognize objects, including ones it has not seen, by their acoustic response[1]. Because the signal is generated by the robot's own action on the object, contact audio is an *active* sense: the manipulator chooses the excitation, much as a person shakes a box to guess its contents.

This places contact audio beside vision-based tactile sensing as a second perception layer that begins at the moment of contact. Where an optical tactile fingertip such as GelSight or DIGIT recovers micron surface geometry and shear from a camera under contact[3], a contact microphone recovers material and interior state from contact vibration. The two are complementary at the fingertip itself, and the Institute's contact-layer report treats the tactile half of this pair[4].

5. Learning manipulation from contact sound

Beyond recognition, in-gripper audio has been shown to teach manipulation itself. ManiWAV mounts a microphone at the gripper and learns contact-rich skills directly from the sounds of interaction — the click of a latch, the scrape of a wipe, the moment of contact — capturing information about contact events and surface properties that a camera, occluded by the gripper and the object, cannot see[2]. Related work uses audio to act under occlusion where vision alone stalls[5], and multisensory object datasets now pair vision, touch, and impact sound for the same objects so that models can learn the cross-modal correspondences[10]. Earlier robotics established the base case: a manipulator can recognize objects from the proprioceptive and auditory feedback of shaking and dropping them[11].

A minimal, transparent version of the recognition step is instructive. Summarize a contact event by a fixed-length *timbre fingerprint*: the log energies in a bank of logarithmically spaced frequency bands, mean-removed and normalized, appended with the spectral centroid of Section 3 and the ring-out time. Given a few labeled example taps per class, store each class's mean fingerprint as a prototype μ_c and classify a new event x by the nearest prototype, reporting a confidence from a softmax over negative distances,

$$p(c \mid x) \;=\; \frac{\exp\big(-\|x-\mu_c\|/T\big)}{\sum_{c'}\exp\big(-\|x-\mu_{c'}\|/T\big)}.$$

This is deliberately the simplest possible learner, the legible counterpart of the learned audio encoders in the systems above: a few examples, a distance, a vote. It is enough to separate a bright metallic ring from a dull wooden thud, which is the whole of the material-identification claim at small scale.

6. Complementarity is coverage, not redundancy

The usual case for fusing sensors is that several noisy measurements of the same quantity average toward the truth. That case is real but it is the weaker one, because it assumes every sense measures the same thing. The stronger case is that different senses are the sole reporters of different properties, and that fusion is therefore about coverage of the property space, not redundancy on a shared axis.

Make this precise. Let an object carry independent latent attributes — say shape, material, fill, and texture — each a categorical variable. Each sense s produces, for each attribute a it can observe, a reading with a known confusion structure, so that its likelihood over that attribute is a vector $L_s(a)$. Under conditional independence of the senses given the object, the fused posterior over an attribute factorizes into a product,

$$p(a \mid \text{readings}) \;\propto\; p(a) \prod_{s\in\mathcal{S}(a)} L_s(a),$$

where $\mathcal{S}(a)$ is the set of senses that actually sense attribute a. The content of the argument is entirely in that set. Assign each sense an acuity per attribute — high, low, or none — reflecting its real strengths: vision resolves shape, contact audio resolves material and fill, touch resolves texture (Table 1). A sense with acuity *none* on an attribute contributes a flat likelihood and drops out of the product; it is not a weak witness but a non-witness.

Table 1. Per-attribute acuity assumed for three senses. Each is the primary — often the sole — witness to one property. Fill has a single witness: sound.

SENSE	SHAPE	MATERIAL	FILL / HOLLOWNESS	TEXTURE
Vision (camera)	high	low	—	low
Contact audio (tap)	—	high	high	low
Touch (fingertip)	low	—	—	high

Two consequences follow, and both are demonstrable by direct computation on the model. First, the fused posterior is sharp on *every* attribute at once, because for each attribute at least one sense is a high-acuity witness, even though each sense alone is confident on one attribute and near-uniform on the rest. Fusion here does not reduce variance on a shared estimate; it assembles a complete estimate from partial, non-overlapping competences. Second, and more pointedly, the column for fill has a single entry. Remove the acoustic channel and the product over $\mathcal{S}(\text{fill})$ becomes empty: the posterior over fill returns to its uniform prior, and identification of that attribute falls to chance, while shape, material insofar as it survives elsewhere, and texture are untouched. No amount of additional visual or tactile resolution recovers it, because the quantity was never in those signals. This is the formal statement of the paper's thesis: *for a hidden property, sound is not redundant with the other senses; it is their only cover.*

> A robot that cannot hear cannot perceive whether a sealed container is full — not poorly, but not at all, at any camera resolution. The missing capability is not accuracy on a known axis; it is an axis.

7. Three transparent instruments

Each argument above is paired with a small, on-device interactive that a reader can run in a browser, built so the mechanism is visible rather than hidden in a trained network. They are didactic constructions, not experiments, and each was verified numerically before it was built.

Complementary Senses implements the model of Section 6 exactly. It samples an object, draws each sense's noisy reading from its confusion structure, and displays the per-attribute posteriors for each sense alone and fused. Over a run of two hundred objects the fused estimate identifies every attribute at high accuracy while each single sense is at chance off its specialty; switching the acoustic channel off collapses the fill row to the chance line, live. *Teach the Ears* runs the fingerprint classifier of Section 5 on the reader's own microphone: tap two objects, teach it a few examples of each, and it names the next tap, with the class prototypes shown as spectra. In synthetic verification it separated two contact timbres at full accuracy from five examples each; on real objects it is honest few-shot, best with distinct sounds at a steady distance. *What the eye can't see* renders the resonance physics of Section 3: an opaque, sealed bottle that looks identical at every fill beside a cross-section that shows the level, with the resonant pitch computed and synthesized as the fill is dragged, in either excitation mode. All three run with no server, no installation, and no data leaving the device.

8. Position and outlook

Contact audio completes a gradient of perceptual reach that the Institute's sensing work traces from the outside in. Distributed remote sensing resolves a volume across the room and, through radio, behind walls[12]; vision-based touch resolves the surface at the moment of contact[4]; contact audio resolves the interior, the material and the hidden state that lie past the surface. Read as a sequence — across the room, at the surface, inside the object — the three name successive depths a single agent must perceive to act competently in the physical world, and they are complementary in the strong sense of Section 6: each is the sole witness to a band of properties the others cannot reach.

The open problems are concrete. Contact audio is an active sense, so the choice of excitation — where and how hard to tap — is itself a perception policy to be learned, not a fixed preprocessing step. The confusion structures assumed in Table 1 are, in real systems, learned from contact data and vary with the object and the grasp; a calibrated, per-object account of when sound is and is not the sole witness would sharpen the coverage argument from a qualitative claim into a design rule. And the fusion itself, written here as a product of independent likelihoods, must contend with correlated failures — a loud environment degrades the acoustic channel and a dropped object degrades all of them at once — which the conditional-independence assumption does not model. None of these blunts the central point. For the properties that live inside things, a robot perceives them by hearing them, or not at all.

References

1. J. Liu, M. Yu, et al. SonicSense: object perception from in-hand acoustic vibration. arXiv:2503.00593, 2025 (Duke University).
2. Z. Liu, C. Chi, E. Cousineau, et al. ManiWAV: learning robot manipulation from in-the-wild audio-visual data. arXiv:2406.19464, 2024 (Stanford / TRI); CoRL 2024.
3. M. Lambeta, P.-W. Chou, S. Tian, et al. DIGIT: a low-cost compact high-resolution tactile sensor. arXiv:2005.14679, 2020; DIGIT 360, arXiv:2411.02479, 2024.
4. D. J. Charlot. The contact layer: vision-based tactile perception where line-of-sight ends. Technical Report TR-2026-09, Institute for Physical AI @ BMI, 2026.
5. A. Du, O. Lee, S. Levine, et al. Play it by ear: learning skills amidst occlusion with audio-visual reinforcement learning. Robotics: Science and Systems (RSS), 2022; arXiv:2205.14850.
6. T. D. Rossing, N. H. Fletcher. Principles of Vibration and Sound, 2nd ed. Springer, 2004.
7. L. E. Kinsler, A. R. Frey, A. B. Coppens, J. V. Sanders. Fundamentals of Acoustics, 4th ed. Wiley, 2000 (Helmholtz resonator, §10).
8. A. P. French. In Vino Veritas: a study of wineglass acoustics. American Journal of Physics 51(8):688–694, 1983.
9. S. Thrun, W. Burgard, D. Fox. Probabilistic Robotics. MIT Press, 2005 (Bayesian sensor fusion).
10. R. Gao, Y. Chang, S. Krishna, et al. ObjectFolder 2.0: a multisensory object dataset for Sim2Real transfer. IEEE/CVF CVPR, 2022; arXiv:2109.07991.
11. J. Sinapov, T. Bergquist, C. Schenck, et al. Interactive object recognition using proprioceptive and auditory feedback. International Journal of Robotics Research 30(10):1250–1262, 2011.
12. D. J. Charlot. OmniSense: perception as a projected volume. Technical Report TR-2026-08, Institute for Physical AI @ BMI, 2026.

Institute for Physical AI @ BMI · The Charlot Lab
Technical Report TR-2026-22 · Survey / Review · Preprint v1 · 21 July 2026
Interactive companions: /research/charlot-lab/topic/acoustic

Preparation: drafted with AI assistance and reviewed by the author.
No new experimental measurements; companions are numerically-verified didactic models.