2025 International Automatic Control Conference (CACS 2025)
Hsinchu, Taiwan, November 5–8, 2025

MicCheck: Repurposing Off-the-Shelf Pin Microphones for Easy and Low-Cost Contact Sensing

Steven Oh*, Tai Inui*, Magdeline Kuan*, Jia-Yeu Lin, *IEEE Member*

***Abstract*—Robotic manipulation tasks are contact-rich, yet most imitation learning (IL) approaches rely primarily on vision, which struggles to capture stiffness, roughness, slip, and other fine interaction cues. Tactile signals can address this gap, but existing sensors often require expensive, delicate, or integration-heavy hardware. In this work, we introduce MicCheck, a plug-and-play acoustic sensing approach that repurposes an off-the-shelf Bluetooth pin microphone as a low-cost contact sensor. The microphone clips into a 3D-printed gripper insert and streams audio via a standard USB receiver, requiring no custom electronics or drivers. Despite its simplicity, the microphone provides signals informative enough for both perception and control. In material classification, it achieves 92.9% accuracy on a 10-class benchmark across four interaction types (tap, knock, slow press, drag). For manipulation, integrating pin microphone into an IL pipeline with open source hardware improves the success rate on picking and pouring task from 0.40 to 0.80 and enables reliable execution of contact-rich skills such as unplugging and sound-based sorting. Compared with high-resolution tactile sensors, pin microphones trade spatial detail for cost and ease of integration, offering a practical pathway for deploying acoustic contact sensing in low-cost robot setups.**

***Index Terms*—Acoustic sensing, imitation learning, low-cost hardware.**

I. Introduction

Imitation learning (IL) has advanced robot manipulation substantially, yet many everyday skills remain contact–rich: task success often hinges on cues that are difficult to perceive with vision alone (e.g., stiffness, roughness, damping, incipient slip, and micro-impacts at the contact interface). Tactile and acoustic feedback can complement vision by sensing contact events directly and by improving robustness under occlusion or challenging illumination. However, practical adoption of tactile sensing faces a deployment gap: higher-performance solutions (e.g., vision-based tactile sensors, custom contact microphones, piezoelectric arrays) tend to be costly, fragile, or integration-intensive (amplifiers, drivers, bespoke software), limiting their use outside well-equipped laboratories. Notably, many manipulation tasks do not require ultra-fine spatial resolution; rather, they benefit from reliable, timely signals that separate no-contact from meaningful contact, discriminate broad material classes, and flag events such as slip or impact. In such regimes, a simpler, lower-cost, and easier-to-integrate sensor can be a reasonable trade-off. We propose *MicCheck*, a plug-and-play approach that repurposes an off-the-shelf Bluetooth pin microphone as a low-cost contact sensor. The microphone clips into a 3D-printed gripper insert and operates out of the box—no custom electronics or drivers—while the stock foam provides compliance and robust acoustic coupling. Despite its simplicity, MicCheck yields signals that are informative for both perception and control: in material classification, it attains 92.9% window-level accuracy on a 10-class benchmark, and in manipulation, incorporating audio into an IL policy improves bottle-cap pouring success from 0.40 (vision only) to 0.80 (vision+audio) and supports two additional contact-rich tasks (texture sorting and high-friction unplugging). These results indicate that low-cost acoustic sensing can provide useful contact awareness, though it does not replace higher-resolution tactile sensors for fine spatial discrimination.

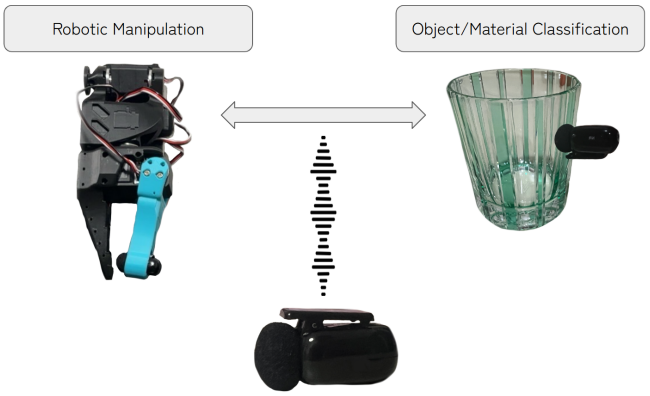

Fig. 1. Overview of MicCheck. We repurpose low-cost pin microphones for contact sensing. We demonstrate this through two experiments: robotic manipulation and object classification.

Our contributions are as follows (Fig. 1): (i) A minimal, low-cost acoustic contact sensor using an unmodified consumer microphone and a simple mechanical mount; (ii) integration into classification and an IL pipeline showing consumer microphone complements imitation learning and is beneficial to contact-rich tasks.

In this paper, Sec. II reviews tactile and acoustic sensing and positions our work. Sec. III details hardware, signal processing, and the learning setup. Sec. IV presents material classification and real-world manipulation results with ablations and discussion. Sec. VI concludes, and Sec. V outlines limitations and future directions.

*Equal contribution.

Authors from this paper are with the Dep. of Modern Mechanical Engineering, Waseda University, Tokyo, Japan.
Corresponding author: Steven Oh (email: oh.steven@fuji.waseda.jp)

II. Related Work

A. Passive Tactile Sensing

Passive tactile sensing methods measure interaction signals without injecting external stimuli. Traditional approaches rely on resistive, capacitive, or force–torque sensors, while more recent systems leverage vision-based tactile designs. Notable examples include the TacTip family of optical biomimetic fingertips [1], GelSight for high-resolution geometry and force estimation [2], and derivatives such as GelSlim [3], DIGIT [4], and OmniTact [5]. Magnetic-based sensors such as ReSkin [6] offer scalable, low-cost tactile skins. Passive acoustic sensing has also been explored, e.g., SonicSense [7], which embeds contact microphones into a multi-fingered hand to capture vibrations for object recognition and material classification. These methods demonstrate that passive signals can encode rich contact information, though often at the cost of custom fabrication or integration.

B. Active Sensing

In contrast, active tactile sensing injects energy—through motion or actuation—into the environment and interprets the response. Active haptic perception surveys [8], [9] highlight how active strategies enable disambiguation of materials and shapes. Lepora et al. demonstrated exploratory tactile servoing using TacTip [10], while more recent work formalizes servoing with pose and shear features [11]. Martinez-Hernandez et al. [12] applied active sensorimotor control for autonomous tactile exploration, and Shahidzadeh et al. [13] combined reinforcement learning with active tactile exploration for shape inference. Within acoustics, Lu and Culbertson [14] demonstrated active acoustic sensing for grasp state estimation, while VibeCheck [15] achieved robust peg-in-hole insertion using only an emitter–receiver acoustic pair. Active methods thus provide controllable, discriminative signals but increase hardware complexity.

C. Acoustic Sensing in Robot Learning

Acoustic signals have recently been explored as a complementary modality for robot learning in contact-rich tasks. Some studies use audio–visual pretraining to improve generalization in low-data regimes [16], while others show that incorporating audio with vision enables better adaptation to texture changes, slip events, and hidden object states [17]. Multimodal systems that fuse vision, touch, and audio also improve performance on tasks such as dense packing and pouring by combining global, temporal, and local cues [18].

Beyond passive use, active acoustic sensing has been applied to infer material properties and grasp states through wave transmission [14], and recent work shows that audio-only feedback can support robust imitation-learned peg-in-hole insertion [15]. Together, these results highlight how acoustic cues capture contact events that are hard to perceive visually or tactually. Our work follows this trend but emphasizes accessibility, using an off-the-shelf microphone as a plug-and-play solution for perception and imitation learning.

III. Method

A. Experimental Setup

We used an existing wireless pin microphone (BOYA mini, model *mini-14*, 2.4 GHz) as acoustic sensor. The clip-on transmitter is press-fit into the gripper mount so that its foam pad serves as the contact interface; the included USB-C receiver (dongle) plugs into the host PC and is enumerated as a standard audio input (Fig. 3). To suppress ambient and motor noise during motion, we enable the microphone's built-in noise-cancellation. Audio is recorded at 48 kHz/16-bit; we use a single transmitter in all experiments.

B. Material Classification

1) Data Collection and Signal Processing: We recorded contact sounds using the pin microphone mounted on the gripper's foam face while interacting with 9 everyday objects spanning rigid solids (plastic lid, glass cup, ceramic mug, steel tumbler, wooden table) and compliant/textured surfaces (leather case, plush toy, notebook), plus a *blank* (no-contact) condition. Speed and normal force were varied across trials.

We collected four interaction types with the selected objects:

- **Tap**: a light, brief touch followed by immediate release.
- **Knock**: a firmer impact that tends to excite audible resonance.
- **Slow press**: a gradual press and hold with mainly normal contact.
- **Drag**: continuous sliding while maintaining contact.

Audio is converted to Mel-spectrogram features. Long recordings are segmented into fixed 1 s windows without overlap; each window is mapped to a log-magnitude Mel spectrogram. This configuration captures resonant peaks, spectral envelopes, and decay characteristics informative for material and structure. Windows with no contact form the *blank* class to avoid forced guesses in the network.

2) Model: We evaluate a compact 2D CNN on single-channel Mel spectrograms (Fig. 2). The network comprises three Conv–BN–ReLU blocks, global adaptive average pooling, and a linear classifier. Windowed examples are split 8:2 (train/validation) with stratification by object class. Models are trained with cross-entropy using Adam (lr $= 3\times10^{-4}$, batch size $= 32$) for 2000 epochs; the best checkpoint is selected by highest validation accuracy.

C. Robot Hardware and Teleoperation Setup

We integrate a commercial pin microphone with the LeRobot SO101 platform [19]. The robot gripper is redesigned to accept the microphone via its built-in clip; the mounting hole is dimensioned for a tight press-fit so the unit can be inserted/removed without additional fasteners. The microphone is oriented perpendicular to the gripper such that its foam pad becomes the primary contact surface on that side during interaction. This foam provides (i) compliant contact for stable grasping and (ii) effective acoustic coupling during object contact, enabling seamless audio capture without modifying the microphone form factor. We enable the device's built-in

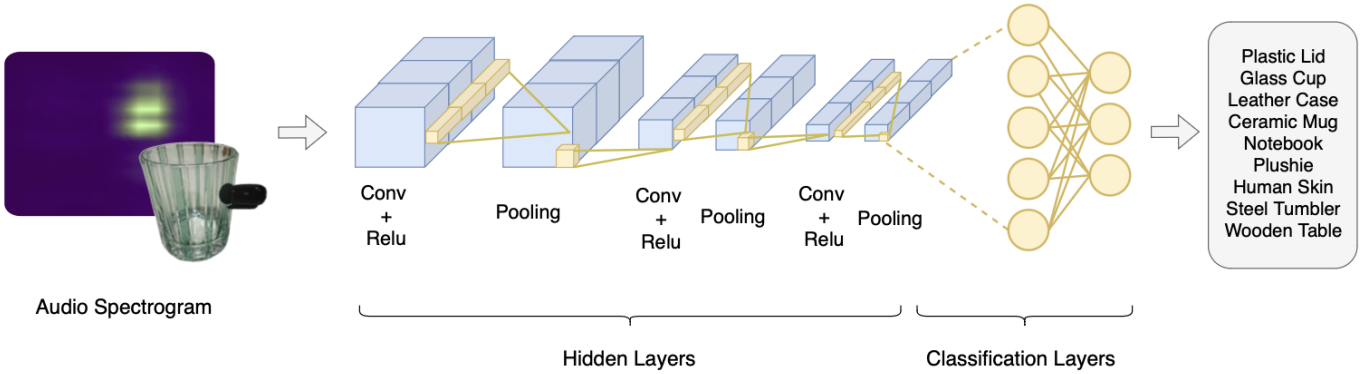

Fig. 2. Architecture of the contact-based object classification model. Single-channel Mel spectrograms from 4 types of mic–object interactions (tap, knock, slow, drag) on 9 objects plus a "blank" no-contact class are fed into a compact 2D CNN comprising three Conv–BN–ReLU blocks, followed by global (adaptive) average pooling and a linear classifier. Models were trained with an 8:2 train/validation split (stratified by class) using cross-entropy loss and the Adam optimizer (learning rate 3×10^{-4}, batch size 32) for 2000 epochs, with the best checkpoint selected by highest validation accuracy. The blank class in the softmax serves as a rejection threshold for low-evidence windows.

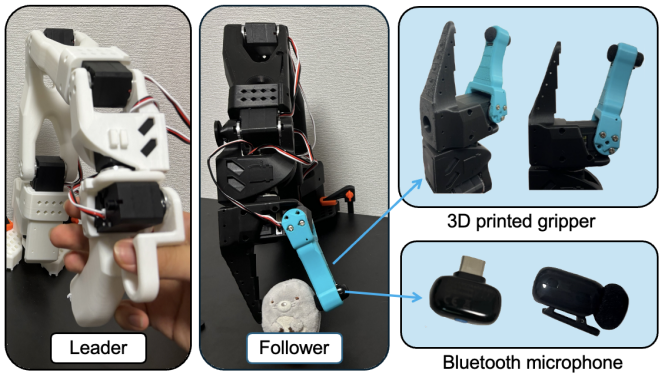

Fig. 3. Teleoperation setup. We employ the Lerobot SO-101 teleoperation setup with a modified gripper. A commonly found commercial bluetooth microphone is embedded onto the gripper. The microphone is connected to a PC via a wireless USB retriever.

noise-cancellation feature to reduce ambient and motor noise, particularly during motion.

For data collection, we use a teleoperation setup with a leader–follower configuration (Fig. 3). Joint states from the leader are streamed to the follower for, a strategy shown to be effective for collecting high-quality demonstrations for robot learning [20]. For training, each dataset consisted of 20 demonstrations per task, collected via teleoperation with synchronized RGB, proprioception, and audio. Compared to vision-only pipelines, the extra audio stream introduced negligible overhead. At inference, the multimodal pipeline ran at 50 Hz, bottlenecked by the camera framerate.

D. Imitation Learning

To capture fine-grained contact dynamics, the audio stream is segmented into 0.2 s frames at 30 Hz (a new frame every 0.04 s; 80% overlap). Each frame is converted into a Mel spectrogram with n_{mels}=32. Frequencies above 0.3× the Nyquist rate are amplified by a factor of 2.0 to emphasize sharp impact sounds. Compared to material classification, we adopt shorter and more overlapping windows here to ensure transient acoustic events remain temporally aligned with proprioceptive and visual signals.

TABLE I
ACT TRAINING PARAMETERS

Parameter	Default
Chunk size	100
Backbone	ResNet-18
Pretrained	ImageNet-1K (ResNet18)
Pre-norm	False
Model dim	512
Heads	8
FFN dim	3200
Activation	ReLU
Enc. layers	4
Dec. layers	1
VAE	True
Latent dim	32
VAE enc. layers	4
Dropout	0.1
KL weight	10.0
Learning rate	1×10^{-5}
Weight decay	1×10^{-4}
Backbone LR	1×10^{-5}

We train an Action Chunking with Transformers (ACT) policy [21] that predicts fixed-length action sequences conditioned on recent multimodal observations (Fig. 4). ACT is an imitation learning approach that mitigates compounding error in sequential control by predicting a chunk of future actions at each step, rather than a single next action. This chunked prediction reduces the effective task horizon and improves the smoothness and robustness of the learned behavior.

In our implementation, observations include (i) RGB images from a stationary camera, (ii) the most recent audio spectrogram frame, and (iii) robot proprioceptive states. These modalities are fused into a unified embedding and processed by a transformer encoder–decoder that outputs target joint positions for the next H timesteps. The model is trained for 100k steps using the hyperparameters listed in Tab. I.

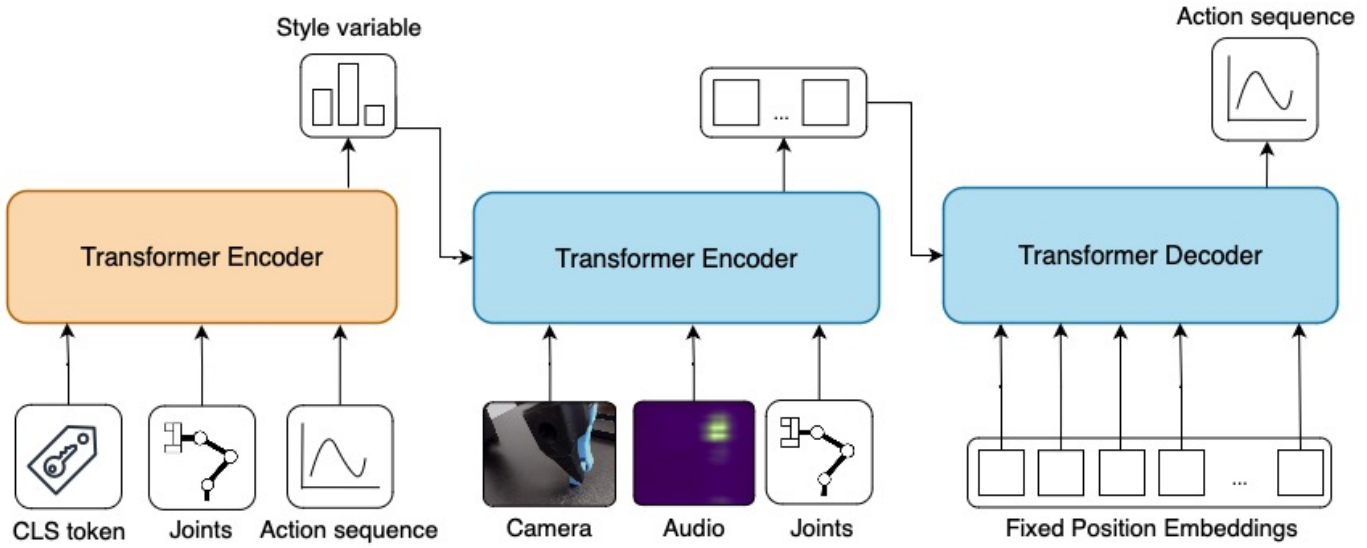

Fig. 4. Action Chunking with Transformers (ACT) architecture. Training uses a conditional variational autoencoder: a transformer episode/style encoder produces a latent $\mathbf{z}$ and a transformer encoder–decoder predicts a chunk of future actions conditioned on observations and $\mathbf{z}$. At inference, the transformer encoder is omitted to generate actions in fixed-size chunks.

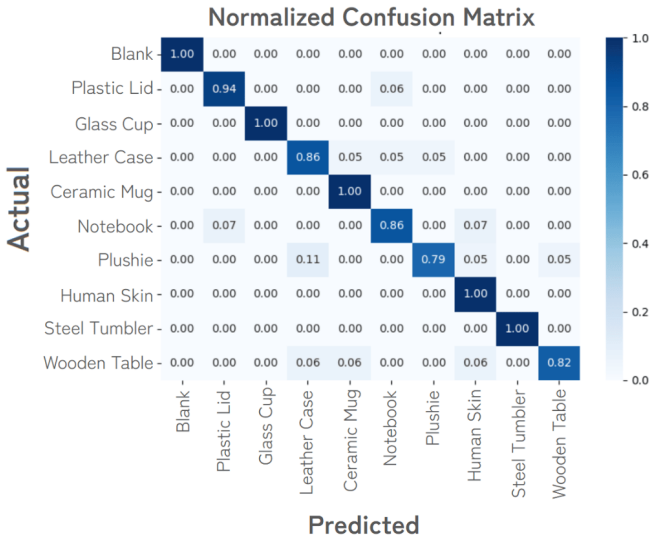

Fig. 5. Normalized confusion matrix for the 10-class (9 objects + "blank") material classification task. The model shows strong diagonal dominance, with perfect accuracy for the blank class, glass cup, human skin, and steel tumbler. Most confusions occur between acoustically similar soft materials (e.g., plushie vs. leather case, notebook vs. leather case), reflecting challenges in distinguishing objects with overlapping frequency responses.

IV. Results

A. Material Classification

We obtain 92.9% window-level classification accuracy on the 9 objects plus a *blank* class. The confusion matrix in Fig. 5 shows strong diagonal dominance, with perfect separation for *blank*, glass cup, ceramic mug, human skin, and steel tumbler, and a high score for plastic lid (0.94). Inaccuracies concentrate in the softer/textured group: plushie is sometimes predicted as leather (0.11) and, to a lesser extent, notebook; leather and notebook also show small mutual bleed while each remains at 0.86 on the diagonal. Wooden table is somewhat less distinct (0.82 on-diagonal), with small leaks into nearby rigid classes (e.g., 0.06 to ceramic). Cross-family errors are rare (rigid items are seldom mistaken for soft ones), and the clean "blank" row/column indicates reliable no-contact rejection. Overall, errors arise mainly among classes with acoustically similar signatures (soft materials and wood) rather than across clearly different materials.

These results indicate that a single, low-cost pin microphone can achieve material discrimination with reasonably high accuracy across a range of object types and interactions. Most misclassifications occur between acoustically similar soft or textured objects, suggesting that the sensor captures the dominant frequency features but may struggle with fine-grained distinctions. The reliable rejection of the *blank* class shows that the system can consistently separate contact from no-contact events, an essential property for integration into robot control pipelines. Overall, the analysis suggests that low-cost pin microphone provides a simple and effective way to add contact awareness across materials and contact types, without needing the resolution of more advanced tactile sensors.

B. Imitation Learning in Real-World Tasks

We first quantify the contribution of audio by ablating sensory inputs on a picking and pouring task (see Fig. 6 a(i)). We use the same dataset but removing the audio tokens from the input at training. As summarized in Table II, success improves from 0.40 with vision only to 0.80 with vision+audio over 10 roll-outs per setting. Fig. 6b illustrates this qualitative difference: the vision-only policy often slips and fails to

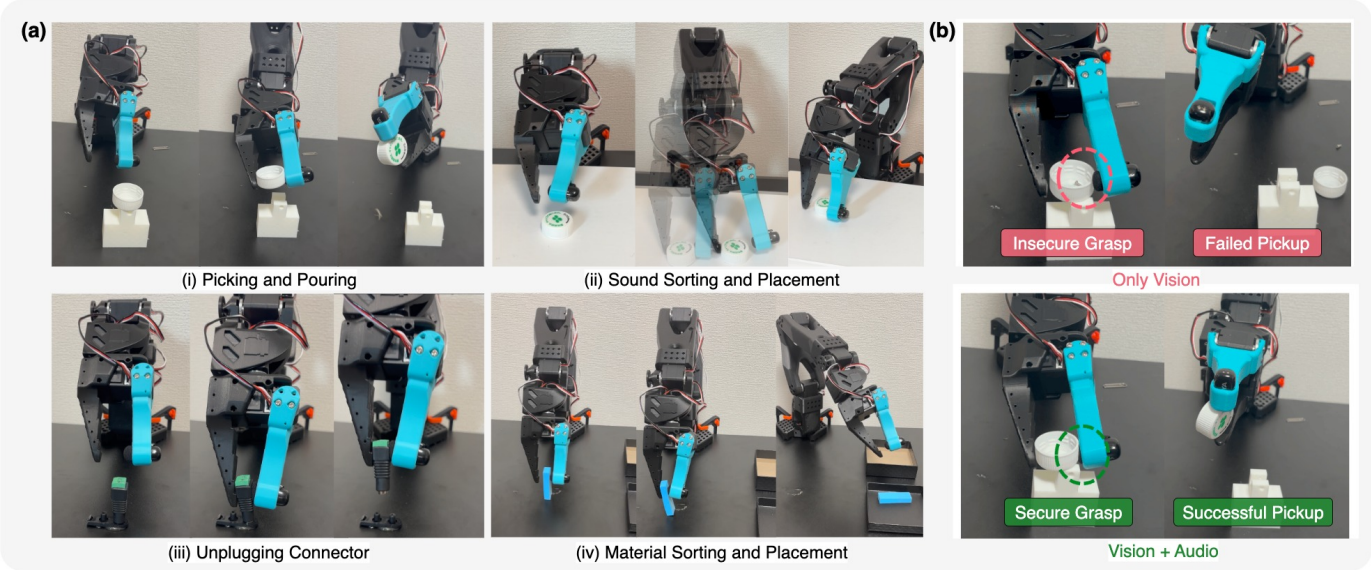

Fig. 6. (a) Demonstration of four manipulation tasks: (i) Picking and pouring — the robot picks up a plastic bottle cap and pours a metallic screw inside; (ii) Sound-based sorting and placement — the robot shakes the cap and sorts it left or right depending on whether a metallic screw is detected acoustically; (iii) Unplugging connector — removing a tightly seated connector that requires high frictional force; and (iv) Material sorting and placement — distinguishing between bare sticky-note pads and ones covered with a plastic film, and sorting them into separate containers. (b) Ablation study on sensing modality tested on task (a-i). With vision-only input, the policy often results in insecure grasps and dropped caps due to the cap's deformability. Incorporating microphone (vision + audio) feedback enables the policy to detect contact states acoustically, achieving more secure and consistent grasps.

TABLE II
REAL WORLD TASK PERFORMANCE

Task	**Condition**	**Input modality**	**Success rate**
(i) Picking and Pouring	—	Vision only	0.40
	—	Vision + Audio	0.80
(ii) Sound Sorting	Has Sound	Vision + Audio	0.70
	No Sound	Vision + Audio	0.60
(iii) Unplugging Connector	—	Vision + Audio	1.00
(iv) Material Sorting	Plastic End	Vision + Audio	0.70
	Normal End	Vision + Audio	0.40

complete the pour, whereas adding audio yields a stable grasp and consistent rotation aligned with demonstrations.

To showcase versatility beyond this ablation, we evaluate three additional contact-rich tasks (Fig. 6). Table II reports their success: unplugging achieves 1.00; sorting by sound reaches 0.70 (has object) and 0.60 (no object); and sorting by texture achieves 0.70 (plastic end) and 0.40 (normal end). Each policy was trained on 20 demonstrations per task (10 per condition for binary tasks) and is evaluated with 10 roll-outs per condition.

Failures primarily stem from insufficient contact leading to aborted attempts, or from incorrect sorting decisions caused by ambient or motor-induced noise. Overall, audio cues proved most reliable when contact interactions generated salient and repeatable sound signatures (e.g., during unplugging). In addition to providing tactile-like feedback through sound, the soft foam surrounding the pin microphone also functioned as a compliant contact interface, improving grip stability and sound coupling during manipulation.

These results show the practicality and value of incorporating a simple commercial microphone for robot learning. Beyond providing complementary acoustic information that vision alone cannot capture, the microphone itself serves as a low-cost and easily integrable sensor that inherently adds compliance through its soft foam housing. Such simplicity lowers the barrier for deploying multimodal perception in everyday robotic systems, making it feasible to scale imitation learning beyond specialized research settings.

V. LIMITATIONS AND FUTURE WORK

While our results show that an off-the-shelf pin microphone can provide useful contact information, several limitations remain. The current system relies on a single consumer-grade microphone with wireless audio, which introduces compression artifacts, latency, and occasional noise contamination. Our experiments also focused on a modest set of objects and interactions, meaning results may not fully generalize to broader manipulation tasks.

Future work should therefore pursue a more rigorous evaluation. This includes comprehensive ablation studies to isolate the role of audio relative to vision and proprioception, including different encoding methods, and expanded benchmarks with larger, more diverse objects and contact types. Furthermore, noise reduction by using pretrained dataset or learning-based approach could improve noise to signal ratio, which needs to be further benchmarked against existing tactile sensors. Additionally, multi-microphone configurations and improved placement strategies using other low-cost sensors

could further extend sensing range beyond single-point contact.

VI. Conclusion

This work explored a minimal and accessible approach to contact sensing for robot learning by repurposing an off-the-shelf Bluetooth pin microphone as an acoustic tactile sensor. We demonstrated that such a simple, low-cost device—integrated via a 3D-printed mount and used without any custom electronics—can produce informative signals for both perception and control. Despite its hardware simplicity, the system achieved high material classification accuracy and improved the robustness of imitation-learned manipulation policies in contact-rich settings. While the performance does not match that of high-resolution tactile sensors, our findings suggest that audio can serve as a lightweight complementary modality to vision and proprioception, offering meaningful cues about contact state, material type, and interaction dynamics. This trade-off between fidelity and accessibility highlights a promising direction for scaling multimodal robot learning beyond specialized laboratory setups.

Overall, we show that commodity microphones, when thoughtfully integrated, can extend the sensory capabilities of everyday robots. We view this not as a replacement for tactile sensors, but as a practical step toward democratizing multimodal sensing—enabling broader experimentation, reproducibility, and adoption of contact-aware learning systems in the robotics community.

Acknowledgment

This work was supported by the JSPS Grant-in-Aid for Early-Career Scientists [23K12755].

References

[1] B. Ward-Cherrier, N. Pestell, L. Cramphorn *et al.*, "The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies," *Soft Robotics*, vol. 5, no. 2, pp. 216–227, 2018.

[2] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[3] E. Donlon, S. Dong, M. Liu, J. Li, E. H. Adelson, and A. Rodriguez, "Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger," in *Proc. IEEE ICRA*, 2018.

[4] M. Lambeta *et al.*, "Digit: A novel design for a low-cost, compact, high-resolution tactile sensor," arXiv:2005.14679, 2020. [Online]. Available: https://ai.meta.com/research/publications/digit-a-novel-design-for-a-low-cost-compact-high-resolution-tactile-sensor-with-application-to-in-hand-manipulation/

[5] A. Padmanabha, F. Ebert, S. Tian, R. Calandra, C. Finn, and S. Levine, "Omnitact: A multi-directional high resolution touch sensor," arXiv:2003.06965, 2020.

[6] R. Bhirangi, T. Hellebrekers, C. Majidi, and A. Gupta, "Reskin: Versatile, replaceable, lasting tactile skins," arXiv:2111.00071, 2021.

[7] J. Liu and B. Chen, "Sonicsense: Object perception from in-hand acoustic vibration," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=CpXiqz6qf4

[8] L. Seminara, P. Gastaldo, S. J. Watt, K. F. Valyear, F. Zuher, and F. Mastrogiovanni, "Active haptic perception in robots: A review," *Frontiers in Neurorobotics*, vol. 13, p. 53, 2019.

[9] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos, "Revisiting active perception," *Autonomous Robots*, vol. 42, pp. 177–196, 2018.

[10] N. F. Lepora, K. Aquilina, and L. P. Cramphorn, "Exploratory tactile servoing with active touch," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 1156–1163, 2017.

[11] N. F. Lepora, J. Lloyd *et al.*, "Pose-and-shear-based tactile servoing," *The International Journal of Robotics Research*, 2024, preprint: arXiv:2312.08411.

[12] U. Martinez-Hernandez, T. Dodd, T. J. Prescott, and N. F. Lepora, "Active sensorimotor control for tactile exploration," *Robotics and Autonomous Systems*, vol. 87, pp. 15–27, 2017.

[13] A.-H. Shahidzadeh, S. J. Yoo, P. Mantripragada, C. D. Singh, C. Fermüller, and Y. Aloimonos, "Actexplore: Active tactile exploration of unknown objects," 2024. [Online]. Available: https://arxiv.org/abs/2310.08745

[14] S. Lu and H. Culbertson, "Active acoustic sensing for robot manipulation," arXiv:2308.01600, 2023.

[15] K. Zhang, D.-G. Kim, E. T. Chang *et al.*, "Vibecheck: Using active acoustic tactile sensing for contact-rich manipulation," arXiv:2504.15535, 2025.

[16] J. Mejia, V. Dean, T. Hellebrekers, and A. Gupta, "Hearing touch: Audio-visual pretraining for contact-rich manipulation," *arXiv preprint arXiv:2405.08576*, 2024. [Online]. Available: https://arxiv.org/abs/2405.08576

[17] Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, B. Burchfiel, and S. Song, "Maniwav: Learning robot manipulation from in-the-wild audio-visual data," *arXiv preprint arXiv:2406.19464*, 2024. [Online]. Available: https://arxiv.org/abs/2406.19464

[18] Y. Li, R. Zhang, Y. Zhu, and D. Xu, "See, hear, and feel: Smart sensory fusion for robotic manipulation," *arXiv preprint arXiv:2212.03858*, 2022. [Online]. Available: https://arxiv.org/abs/2212.03858

[19] R. Cadene, S. Alibert, A. Soare, Q. Gallouedec, A. Zouitine, S. Palma, P. Kooijmans, M. Aractingi, M. Shukor, D. Aubakirova, M. Russi, F. Capuano, C. Pascal, J. Choghari, J. Moss, and T. Wolf, "Lerobot: State-of-the-art machine learning for real-world robotics in pytorch," https://github.com/huggingface/lerobot, 2024.

[20] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators," 2023.

[21] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," 2023. [Online]. Available: https://arxiv.org/abs/2304.13705