# SonicSense:
# Object Perception from In-Hand Acoustic Vibration

**Jiaxun Liu Boyuan Chen**
Duke University
http://generalroboticslab.com/SonicSense

**Abstract:** We introduce SonicSense, a holistic design of hardware and software to enable rich robot object perception through in-hand acoustic vibration sensing. While previous studies have shown promising results with acoustic sensing for object perception, current solutions are constrained to a handful of objects with simple geometries and homogeneous materials, single-finger sensing, and mixing training and testing on the same objects. SonicSense enables container inventory status differentiation, heterogeneous material prediction, 3D shape reconstruction, and object re-identification from a diverse set of 83 real-world objects. Our system employs a simple but effective heuristic exploration policy to interact with the objects as well as end-to-end learning-based algorithms to fuse vibration signals to infer object properties. Our framework underscores the significance of in-hand acoustic vibration sensing in advancing robot tactile perception.

**Keywords:** Tactile Perception, Object State Estimation, Audio

## 1 Introduction

By shaking a container, we can tell its inventory status from the generated acoustic vibrations, such as the quantity and geometry of the objects inside. Similarly, we can identify the material and geometry of the entire object through multiple tappings. However, despite the significance of acoustic vibrations for tactile perception, equipping robot manipulators with acoustic vibration sensing capability for rich object perception remains difficult [1, 2, 3, 4, 5].

Though previous research has explored placing air microphones near robot platforms to estimate liquid height [6] and pouring amounts [7], classify object materials [8] and categories [9, 10, 11], air microphones mainly capture sound waves transmitted through air, leading to noisy signals with ambient noises. On the other hand, contact microphones only sense the acoustic vibrations caused by physical contact. Past work has studied contact microphones for estimating the amount and flow of granular material [12], object position and category [13], and collectively performing object spatial reasoning for visual reconstruction [14].

Several major challenges remain to advance acoustic vibration sensing for robot object perception. Most current solutions focus on constrained settings with a small number ($N < 5$) of primitive objects [6, 7, 8, 12, 14, 15], homogeneous material composition for each object [8, 12, 15, 16], single-finger testing [15, 16], and training and testing on different contacts but same objects [15, 16]. However, it is not clear whether such testing results can work with noisy and less controlled conditions. In addition, previous computational algorithms mainly utilize small machine learning models [9, 11, 15, 16] with a limited amount of data, which could be difficult to generalize. Moreover, the interaction mechanisms to collect acoustic data with objects rely on human manual movements [14, 15] or replaying pre-defined fixed robot poses [6, 7, 8, 9, 10, 12, 13, 15], making it difficult to scale to a large number of objects.

We present SonicSense (Fig. 1), a holistic design on both hardware and algorithm advancements for object perception through in-hand acoustic vibration sensing. Our design enables effective object

8th Conference on Robot Learning (CoRL 2024), Munich, Germany.

perception abilities that are difficult for previous approaches to achieve altogether on 83 diverse real-world objects, including objects with complex geometry and heterogeneous materials. Our robot is capable of differentiating the inventory status of an occluded container through interactions. In more challenging tasks, through a naive but effective heuristic exploration policy to autonomously collect acoustic vibration characteristics, we can successfully infer material compositions, reconstruct the complete 3D object shape through sparse tapping, and re-identify previous objects base on a set of end-to-end learning-based models by leveraging our large-scale dataset. Moreover, our design is cost-effective ($215.26) and easy to build. Overall, our method presents unique contributions and opens up new opportunities for robot tactile perception.

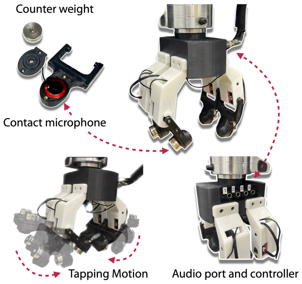

Fig. 1: Our robot hand includes four fingers where each fingertip is equipped with one contact microphone and a counterweight.

## 2 The SonicSense

**Robot Hand Design** Our robot hand (Fig. 1) has four fingers and each finger has one joint with one degree of freedom. At each fingertip, a piezoelectric contact microphone is embedded inside the plastic shell to record acoustic vibration signals while a round counterweight is mounted on the outer shell surface to increase the momentum of the finger motion. We found that the counterweight plays an important role in enabling large striking vibrations during tapping motion.

**Real-World Object Dataset on Acoustics, Material, Shape, and Category** We have developed a dataset with 83 diverse real-world objects shown in Fig. 2. Our dataset covers nine material categories including challenging materials such as foam and fabric, and 22.9% of the objects include more than one material. Our objects cover a variety of geometries, from simple primitives to complex shapes and from smaller objects to larger or longer objects.

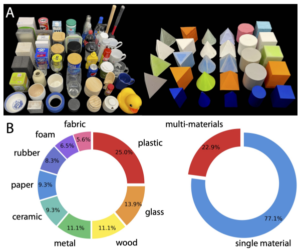

Fig. 2: (A) 54 everyday objects and 29 3D-printed primitive objects with different materials attached to their surfaces. (B) The composition of the nine materials and multi-material vs. single-material objects.

**Heuristic-Based Interaction Policy** We derive a simple but effective heuristic-based tapping motion to collect the acoustic vibration response from all our real-world objects covering variable sizes and geometries. First, due to the unknown shape of the object, the policy will attempt to make contact with the object from high to low heights with a fixed step size, until the first contact event is detected from acoustic signals. Second, from such initial exploration, we can estimate the height and the radius of the object. Finally, the robot will use a grid sampling schedule to make sparse tapping contact with the objects to collect acoustic responses.

## 3 Object Perception from In-Hand Acoustic Vibration

**Material Classification Model and Training** The network shown in Fig. 4(A) takes in the Mel-spectrogram $A_i$ of the i$^{\text{th}}$ sample with label $m_i$ in our interaction data. We train the material classification model $f_{mc}$ to output the corresponding material label category $\hat{m}_i = f_{mr}(A_i)$. We optimize the model with the cross-entropy loss $\mathcal{L}_{mc}(\hat{m}_i, m_i)$.

**Shape Reconstruction Model and Training** As shown in Fig. 4(B), given a contact point cloud $C_i$ of the i$^{\text{th}}$ object collected through our robot hand interaction, we train our shape reconstruction model $g_{sr}$ to generate a dense and complete point cloud of the corresponding object $\hat{P}_i = g_{sr}(C_i)$. We optimize the model with the Chamfer Distance loss $CD(\hat{P}_i, P_i)$.

Because of the challenging nature of this task and the limited amount of real interaction data, we constructed a simulation environment to augment the dataset as shown in Fig. 3. We first pre-train the network with only our synthetic dataset to capture necessary prior knowledge and then gradually reduce the percentage of synthetic data and increase the percentage of real-world data during the training process.

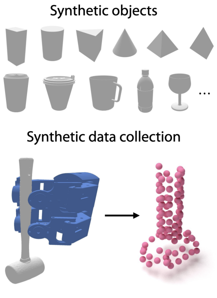

Fig. 3: We conducted synthetic data collection of contact points on a large number of 3D objects in the simulation for data augmentation.

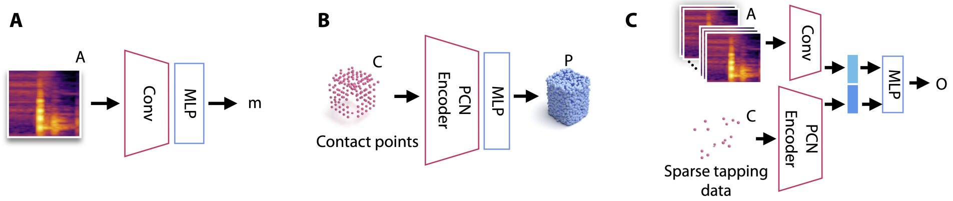

Fig. 4: **The network architectures.** (A) Material classification network. (B) Shape reconstruction network (C) Object re-identification network.

**Object Re-identification Model and Training** When an object has been interacted with by the robot, we aim to have our robot re-identify the object through a set of new tapping interactions. In our object re-identification model, as shown in Fig. 4(C), we input both the collection of a set of fifteen spectrograms $A_i$ and contact point cloud $C_i$ of the i$^{\text{th}}$ object in our interaction data and train the shape reconstruction model $h_{or}$ to predict the corresponding object label $\hat{o}_i = h_{or}(A_i, C_i)$. We optimize the model with the cross-entropy loss $\mathcal{L}_{or}(\hat{o}_i, o_i)$.

## 4 Experimental Results

**Characterizing Basic Sensing Capabilities of SonicSense** To assess whether SonicSense design can capture subtle but informative acoustic vibration signals to reveal object states in challenging scenarios, we conducted two experiments with a focus on differentiating inventory status in containers. We first placed different numbers of dice and then a series of dice with various shapes inside a plastic container. We had the robot hold the container and rotate it forward and backward by 180° around the wrist. In the second experiment, the robot held a bottle with three different initial amounts of water (e.i. $0\,\mathrm{mL}$, $100\,\mathrm{mL}$, $200\,\mathrm{mL}$). We then poured $100\,\mathrm{mL}$ water three separate times into the bottle with a constant flow using a dispenser bottle. Next, the robot held the bottle and performed a horizontal shaking motion with three amounts of water (i.e., $100\,\mathrm{mL}$, $200\,\mathrm{mL}$, and $300\,\mathrm{mL}$).

From the visualization of the captured vibration signals in Fig. 5, we can tell that the signals reflect the spatial and temporal features of different inventory statuses. Quantitatively, we derived twelve interpretable features based on traditional acoustic signal processing, including the root mean square of the signal, spectral centroid, bandwidth, contrast, flatness, roll-off, zero crossing rate, tempogram, poly features, Mel-frequency cepstral coefficients, chroma, and tonnetz, all averaged across time. We then performed an unsupervised nonlinear dimensionality reduction with t-SNE [17] on this 12-dimensional feature vector for all our experiments as shown in Fig. 5. The clear clusters indicate that SonicSense is able to provide informative cues to distinguish not only the numbers and geometries of solid objects but also the continuous and subtle liquid states in a small container.

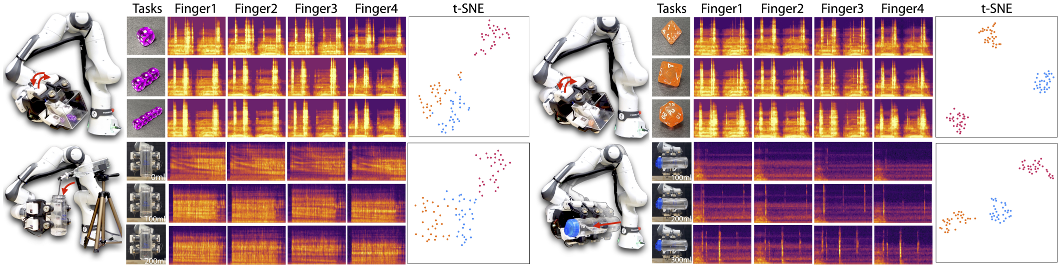

Fig. 5: The Mel-spectrograms show one example of the collected acoustic vibration signal. The t-SNE results are based on 30 trials for each object across all experiments. Different sub-tasks are represented by the three colors.

**Material Classification** As shown in Fig. 6(A), the initial result of our method leads to a 0.523 F1 score. However, many errors stem from outlier-like predictions, even though most predictions in the surrounding object regions remain accurate. Therefore, we propose an iterative refinement procedure assuming that materials are relatively uniform and smooth around local regions. Our iterative algorithm works as follows: for each object, we first filter out the predictions with low occurrence with a threshold $M$ and reassign their labels with the highest occurrence label. For each point, we assign the label based on a majority vote among all its $K$ nearest neighbors. We then repeat this step for $N$ steps. The values of $M$, $K$, and $N$ are selected based on the best validation performance. With this refinement algorithm, our final average F1 score reaches 0.763.

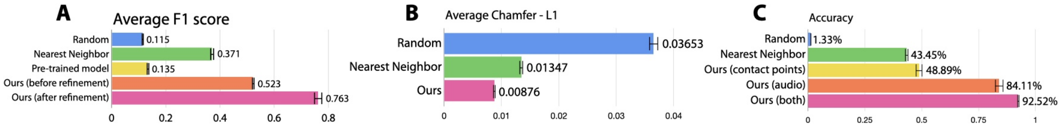

Fig. 6: Quantitative results and baseline comparisons of tasks of (A) material classification, (B) shape reconstruction, and (C) object re-identification. All our experiments are conducted with three different splits to obtain the mean and standard deviation. For the nearest neighbor baseline, we compared each input sample in the test dataset with all samples in the training dataset and selected the label based on the lowest mean square distance for the mel-spectrogram and Chamfer-L1 distance for the contact points.

Our method outperforms both random and nearest neighbor baselines, suggesting that our algorithm generalizes beyond the training set and provides accurate material label prediction on *unseen objects*. Moreover, we experimented with pre-training our model with the recent audio-material dataset [5, 18, 19]. However, we found that this pre-training scheme hurts the performance. The acoustic signals in these datasets were collected with air microphones and noise-controlled experiments. Hence, a large domain gap exists.

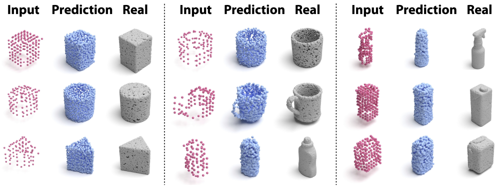

Fig. 7: SonicSense can produce a complete and accurate 3D point cloud of objects from sparse, nonuniform, and noisy contact positions.

**Shape Reconstruction** As shown in Fig. 6(B), through sparse contact tapping points, we obtained an average of $0.00876\text{m}$ Chamfer-L1 distance score. Fig. 7 shows examples of our shape reconstruction results on our testing dataset. Our model greatly outperforms both baselines and shows strong generalization abilities on unknown shape reconstructions. The prediction on objects with primitive shapes generally has near-perfect performance. Additionally, our method exhibits the capability to reconstruct objects with concave geometries. Some failure prediction examples include the nozzle of the spray and the cap of the bottle, due to the limited number of spray objects in our dataset and its complex shape within a small region.

**Object Re-identification** Our model can accurately re-identify the objects with a $92.52\%$ test accuracy while the random baseline and nearest neighbor baseline only gave $1.33\%$ and $43.45\%$ respectively, as shown in Fig. 6(C). By looking into the detailed confusion matrix of our testing results, we can see that smaller objects are generally more difficult due to the limited number of interaction data. We can also observe that the model performs worse when the objects are relatively small, and the materials have similar acoustic properties, such as ceramic and glass.

Additionally, in order to verify the importance of both the shape and material information for this object re-identification task, we conducted two ablation studies to either remove the acoustic vibration input or the tapping point input. The acoustic vibration-only network reaches an accuracy of $84.11\%$, and the tapping point-only network reaches an accuracy of $48.89\%$. Therefore, acoustic information provides a more informative representation of the object and the performance will be further improved by incorporating an additional modality of rough contact positions for object re-identification.

## 5 Conclusion

We have introduced SonicSense, an integrated hardware and software solution to enable rich object perception capabilities with in-hand acoustic vibration for a multi-finger robot hand. Our experimental results demonstrate the versatility and efficacy of our design on a variety of object perception tasks. Our study involves a significantly larger number of real-world objects with complex geometry and heterogeneous materials. Our investigations outline the challenges and necessities of considering real-world noises and robot-specific interactions for robot object perception. Despite these advancements, we see many opportunities to improve our current approach. One immediate future work can consider adapting object tracking as an online object estimation and tracking within the interaction policy to avoid fixing the objects on the table. Future work can also integrate acoustic vibration sensing along with multiple sensing modalities in a higher DoF robot hand for dexterous manipulation tasks.

**Acknowledgments**

This work is supported by ARL STRONG program under awards W911NF2320182 and W911NF2220113, by DARPA FoundSci program under award HR00112490372, and DARPA TIAMAT program under award HR00112490419.

## References

[1] Z. Zhang, Q. Li, Z. Huang, J. Wu, J. Tenenbaum, and B. Freeman. Shape and material from sound. *Advances in Neural Information Processing Systems*, 30, 2017.

[2] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016.

[3] A. Arnab, M. Sapienza, S. Golodetz, J. Valentin, O. Miksik, S. Izadi, and P. Torr. Joint object-material category segmentation from audio-visual cues. *arXiv preprint arXiv:1601.02220*, 2016.

[4] S. Luo, L. Zhu, K. Althoefer, and H. Liu. Knock-knock: acoustic object recognition by using stacked denoising autoencoders. *Neurocomputing*, 267:18–24, 2017.

[5] R. Gao, Y.-Y. Chang, S. Mall, L. Fei-Fei, and J. Wu. Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. *arXiv preprint arXiv:2109.07991*, 2021.

[6] H. Liang, S. Li, X. Ma, N. Hendrich, T. Gerkmann, F. Sun, and J. Zhang. Making sense of audio vibration for liquid height estimation in robotic pouring. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5333–5339. IEEE, 2019.

[7] J. Wilson, A. Sterling, and M. C. Lin. Analyzing liquid pouring sequences via audio-visual neural networks. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7702–7709. IEEE, 2019.

[8] S. Clarke, N. Heravi, M. Rau, R. Gao, J. Wu, D. James, and J. Bohg. Diffimpact: Differentiable rendering and identification of impact sounds. In *Conference on Robot Learning*, pages 662–673. PMLR, 2022.

[9] J. Sinapov, T. Bergquist, C. Schenck, U. Ohiri, S. Griffith, and A. Stoytchev. Interactive object recognition using proprioceptive and auditory feedback. *The International Journal of Robotics Research*, 30(10):1250–1262, 2011.

[10] C. Schenck, J. Sinapov, D. Johnston, and A. Stoytchev. Which object fits best? solving matrix completion tasks with a humanoid robot. *IEEE Transactions on Autonomous Mental Development*, 6(3):226–240, 2014.

[11] T. Nakamura, T. Nagai, and N. Iwahashi. Multimodal object categorization by a robot. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2415–2420. IEEE, 2007.

[12] S. Clarke, T. Rhodes, C. G. Atkeson, and O. Kroemer. Learning audio feedback for estimating amount and flow of granular material. *Proceedings of Machine Learning Research*, 87, 2018.

[13] D. Gandhi, A. Gupta, and L. Pinto. Swoosh! rattle! thump!–actions that sound. *arXiv preprint arXiv:2007.01851*, 2020.

[14] B. Chen, M. Chiquier, H. Lipson, and C. Vondrick. The boombox: Visual reconstruction from acoustic vibrations. In *Conference on Robot Learning*, pages 1067–1077. PMLR, 2022.

[15] V. Wall, G. Zöller, and O. Brock. Passive and active acoustic sensing for soft pneumatic actuators. *The International Journal of Robotics Research*, 42(3):108–122, 2023.

[16] S. Lu and H. Culbertson. Active acoustic sensing for robot manipulation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023.

[17] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[18] R. Gao, Z. Si, Y.-Y. Chang, S. Clarke, J. Bohg, L. Fei-Fei, W. Yuan, and J. Wu. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10598–10608, 2022.

[19] R. Gao, Y. Dou, H. Li, T. Agarwal, J. Bohg, Y. Li, L. Fei-Fei, and J. Wu. The objectfolder benchmark: Multisensory learning with neural and real objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17276–17286, 2023.