2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Acoustic Sensing for Universal Jamming Grippers

Lion Weber*,1,2 Theodor Wienert*,1 Martin Splettstößer*,1,2 Alexander Koenig[1,3] Oliver Brock[1,2,3]

***Abstract*— Universal jamming grippers excel at grasping unknown objects due to their compliant bodies. Traditional tactile sensors can compromise this compliance, reducing grasping performance. We present acoustic sensing as a form of *morphological sensing*, where the gripper's soft body itself becomes the sensor. A speaker and microphone are placed inside the gripper cavity, away from the deformable membrane, fully preserving compliance. Sound propagates through the gripper and object, encoding object properties, which are then reconstructed via machine learning. Our sensor achieves high spatial resolution in sensing object size ($2.6\,\mathrm{mm}$ error) and orientation (0.6° error), remains robust to external noise levels of $80\,\mathrm{dBA}$, and discriminates object materials (up to $100\,\%$ accuracy) and 16 everyday objects ($85.6\,\%$ accuracy). We validate the sensor in a realistic tactile object sorting task, achieving 53 minutes of uninterrupted grasping and sensing, confirming the preserved grasping performance. Finally, we demonstrate that disentangled acoustic representations can be learned, improving robustness to irrelevant acoustic variations.**

I. Introduction

Universal jamming grippers [2] grasp by enveloping an object with a flexible membrane filled with granular media, and then stiffening through jamming. This simple, compliant mechanism lets them conform to a wide variety of unknown objects, making them particularly attractive for grasping in unstructured environments. Despite their versatility, universal jamming grippers remain difficult to sensorize: traditional rigid sensors compromise compliance, reducing grasping performance. Yet, tactile sensing is essential in realistic environments. Acoustic sensing offers a solution that preserves compliance while providing rich tactile feedback.

Acoustic sensing is a form of *morphological sensing* [3], where the gripper's soft body itself functions as the sensor. The key idea is that interactions between the gripper and environment induce measurable changes in the gripper's morphology. Sound propagating through the gripper captures these changes as modulated signals, from which machine learning reconstructs desired sensor measurements.

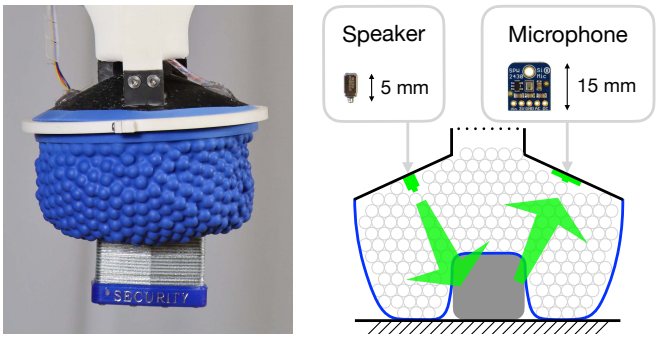

Fig. 1. Left: Universal jamming grippers conform to unknown objects, enabling versatile grasping, but traditional sensors restrict this deformability. Right: We address this with acoustic sensing. A speaker emits sound into the gripper cavity, which propagates through the gripper and object, encoding object properties (size, material, orientation, class, etc.) in the modulated signal shown in green. A microphone then records the signal, and machine learning reconstructs the object state.

In this paper, we explain how to build an acoustic sensor that fully preserves the gripper's compliance (Fig. 1). We then show that a single acoustic sensor can perceive many physical quantities that may otherwise require specialized sensors. The acoustic sensor estimates object *size* with millimeter-scale resolution and remains robust to substantial external acoustic noise. It reliably senses object *materials*, which is impossible for cameras when objects appear visually identical. Our approach also senses the object *orientation* with sub-degree resolution. In a realistic application, we discriminate the *class* of 16 household objects with a high accuracy of $85.6\,\%$ and show the gripper's unaffected grasping performance in a tactile object-sorting task. Finally, we demonstrate that the high-dimensional acoustic data can be transformed into disentangled latent spaces, isolating task-relevant information and improving robustness against irrelevant variations such as object pose.

II. Related Work

We first analyze existing sensing methods for universal jamming grippers, with a focus on their impact on gripper compliance, and then present prior art on acoustic sensing.

A. Sensorized Universal Jamming Grippers

Some prior works on sensorizing universal jamming grippers rely on mounting rigid cameras on or inside the gripper. For example, a small camera can be placed at the center of the gripper membrane [4], but the rigid mount inside the gripper likely worsens compliance. Other efforts relying on cameras [5]–[7] fill the gripper cavity with liquid and glass

* Equal contributions
[1] Robotics and Biology Laboratory, Technische Universität Berlin
[2] Science of Intelligence, Research Cluster of Excellence, Berlin
[3] Robotics Institute Germany

This paper extends our workshop contribution [1] by evaluating sensor resolution across object orientations and providing an extensive discussion of the approach. Portions of the earlier version, including text and figures, are reused or adapted in this manuscript.

This work has been partially supported by the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG) and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 – project no. 390523135; and project no. 405033880.

Project page with videos and code: https://rbo.gitlab-pages.tu-berlin.de/papers/acoustic-jamming-icra26

beads of the same refractive index. Cameras mounted inside the gripper can then see the membrane deformation through the glass beads to estimate the object's shape and pose. A challenge of these approaches is protecting the electronics inside the gripper and other robot parts from liquid leakage. Most of these works [6], [7] require custom membranes with physical markers, which were shown to worsen grasping capabilities [6] and are complex to manufacture compared to using latex balloons as membranes [2].

Other works directly sensorize the gripper membrane with flexible sensor arrays [8], [9]. These sensors can predict proximity, membrane deformation, normal forces [8], and slippage [9]. However, the attached sensor arrays again limit the membrane's compliance, require a complex manufacturing process, and have coarse spatial resolution of 6×6 [8] or 3×3 [9] sensor grids. Our sensor is mounted far away from the membrane and therefore preserves its compliance, is easy to manufacture, and can accurately reconstruct various physical quantities from a single sensor signal.

B. Acoustic Sensing

Industry and robotics research increasingly leverage acoustic signals. In industrial settings, acoustic sensing detects tool wear [10], rail infrastructure failure [11], and motor defects [12]. In robotic manipulation with rigid grippers, acoustic sensing classifies household objects [13], [14], estimates object materials and contact positions [14], [15], and enables peg-in-hole insertion [16]. Acoustic sensing is also applied in soft robotic grasping. Wall et al. [17] present an acoustic sensor integrated into soft robotic fingers, which accurately predicts contact locations, forces, materials, and temperature. They demonstrated that *active* acoustic sensing, where sound is injected rather than passively observed, yields more robust predictions [17]. Other works reconstruct the deformed shape of soft robotic actuators using acoustics [18], [19]. While prior work demonstrated the versatility of acoustic sensing across many robotics domains, it has not yet been applied to universal jamming grippers—despite these grippers being challenging to sensorize due to their deformable nature. To our knowledge, we are the first to bridge this gap.

III. Acoustic Sensing for Universal Jamming Grippers

This section explains how to build and operate the acoustic jamming gripper, showing how the gripper's deformable body, which is normally a challenge for sensorization, turns into an *asset* for morphological sensing.

A. The Gripper Body as a Sensor

Our design goal is to augment the gripper with sensing capabilities while fully preserving its compliance. Acoustic sensing offers a solution: the hardware, a speaker and microphone, can be placed away from the deformable membrane, leaving the compliant morphology untouched (Fig. 1). But acoustic sensing does more than just preserve compliance—it exploits it. The same deformability that lets the gripper conform to objects for a stable grasp also shapes the acoustic

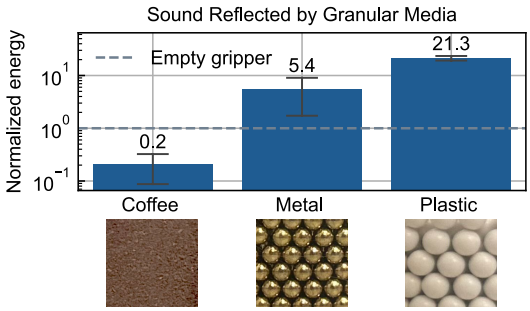

Fig. 2. Different granular media inside the gripper have different sound absorption characteristics. Good sound reflection is necessary to return information about the environment to the microphone. The plot shows the energy of the sound signal as reflected by different granular media inside the gripper. Results are scaled relative to the sound reflection in an empty gripper (i.e., when only air is inside). We use plastic ball bearings as our granular medium because they better reflect the speaker's sound than ground coffee and metal ball bearings.

signal in a way that reveals object properties. Acoustic sensing leverages the same morphology as a resonant structure: morphology enables *grasping*—and it enables *sensing* too. When the soft jamming gripper conforms to an object, the resulting imprint in the cavity encodes information about the object's geometry and pose relative to the gripper. The large contact area between gripper and object facilitates energy transfer *into* the object, revealing intrinsic object properties such as material composition. The microphone captures the reflected sound waves, modulated by both the altered cavity and the object itself. Subsequent computation then extracts object properties embedded in this modulation.

B. Building an Acoustic Jamming Gripper

Our implementation builds on the classic universal jamming gripper design [2]. A latex balloon filled with granular medium is mounted on a rigid conical housing and connected to a vacuum source (Fig. 1). At ambient pressure, the balloon deforms around the object, conforming to its shape. Applying negative pressure jams the granular medium, stiffening the structure and securing the grasp. The gripper design files, all data, and software are available in our code repository.

Installing the electronics on the jamming gripper is simple. We mount a small speaker (Knowles RAB-32063-000) and microphone (Adafruit SPW2430) inside the cavity. Wires are routed through $2\,\mathrm{mm}$ holes in the housing and sealed with hot glue. The speaker and microphone are positioned $7\,\mathrm{cm}$ apart and angled $25°$ toward the gripper's contact region to maximize acoustic transmission and reception while avoiding direct coupling (Fig. 1, right). Importantly, both elements are rigidly fixed to the housing rather than the membrane, ensuring that the deformable body remains unimpeded. The microphone signal is amplified using a small preamplifier (Adafruit PAM8302A) and digitized through an 18-bit analog-to-digital converter at $44.1\,\mathrm{kHz}$ (MAYA44 USB+). The speaker signal is generated via the 20-bit digital-to-analog converter of the same interface. All supporting electronics, including amplification and audio interface, are located outside the gripper. The total cost of the sensing

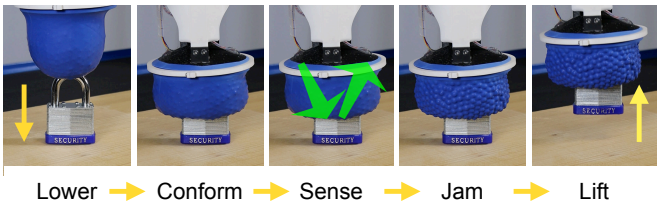

Fig. 3. The sensing and grasping process. We sense after conforming to the object, when the contact area for acoustic energy exchange is large, but before jamming, because the lower air pressure during evacuation reduces sound transmission in the cavity and the vacuum pump generates substantial noise. The supplementary video shows the grasping process.

electronics is below 140 USD. Overall, this demonstrates the setup is low-cost, simple to integrate, and fully compatible with the soft, deformable structure of the gripper.

Acoustic sensing relies on the reflection of sound: the gripper filling must transmit and reflect energy so the microphone can capture signals modulated by the object. Therefore, selecting the right granular medium is crucial as it determines how effectively sound can travel towards the object and back to the microphone. Fig. 2 shows that coffee grounds, commonly used in jamming grippers [2], attenuate more than 80 % of the acoustic energy and are therefore unsuitable for sensing. This result is in line with previous research showing that coffee has strong sound-insulating properties [20]. We compare metallic and plastic ball bearings as alternative fillings. Ball bearings of plastic (6 mm diameter, 0.2 g, Elite Force WA41839) and metal (4.5 mm diameter, 0.36 g, Walther Premium Steel BBs 4.1668-1) reflect the sound better. We chose plastic ball bearings because they provide the best sound reflection, while metal ball bearings are heavy and can short-circuit electronics.

C. Using an Acoustic Jamming Gripper

Fig. 3 shows the proposed sensing and grasping process. First, we loosen the medium and approach the object with an impedance controller on a Franka Panda arm. The gripper conforms to the object until an ATI Mini40 sensor on the gripper mount reaches a contact force of 8 N. We then keep the gripper in position. We allow two seconds for the settling of the granular medium and for venting air that is displaced during the compression of the cavity. The remaining contact force facilitates vibration exchange between the granular medium and the object. The sensing process happens at ambient pressure and not while jamming because the negative pressure reduces sound transmission through air, and the vacuum pump induces vibrations that propagate into the cavity. We play a one-second logarithmic frequency sweep from 20 Hz to 20 kHz on the speaker to emphasize lower frequencies, which tend to travel farther than high frequencies. The microphone records for one second as the sound plays. After reconstructing the object properties from the reflected audio signal, we jam the medium by evacuating the gripper, and lift the object.

We process the one-second microphone recordings following the approach by Wall et al. [17]. We compute a short-time Fourier transform (STFT) with a window size of 2048 to obtain 1025 frequency bins. We sum the STFT output across the time axis and normalize to get our final 1025-dimensional feature vector. This representation encodes the morphological imprint of the object, including cues about object properties. To decode these morphological cues, we employ a three-layer convolutional neural network (CNN) followed by a fully connected layer. This last layer maps to a regression value or a vector of class probabilities via a softmax operation, depending on the task. Our loss is a mean-squared error for regression and cross-entropy for classification. We use an 80:20 train-validation split in a 5-fold cross-validation. All results in this paper report the validation mean and variance from this cross-validation unless indicated otherwise.

Fig. 4. The object set used in this paper (*bottom*: cubes of different sizes for Section IV-A; *left*: plates and spheres of different materials for Section IV-D; *right*: everyday objects from the YCB [21] dataset for Section V)

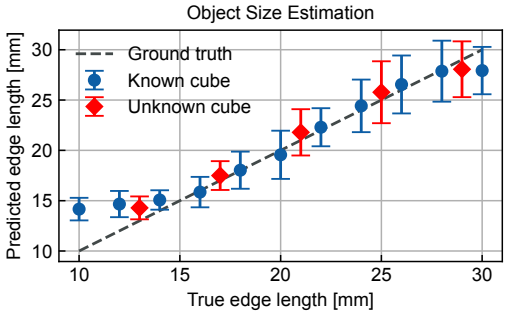

Fig. 5. The acoustic sensor has high spatial resolution: it can accurately predict the size of enclosed objects. Our model predicts the cube edge length with a small root mean square error (RMSE) of 2.7 mm for known objects, 2.4 mm for unknown objects, and an overall error of 2.6 mm. Fig. 4 shows the corresponding cubes.

IV. Sensor Evaluation

This section evaluates the sensor's resolution, robustness to external acoustic noise, and discriminative performance across quantities, including object size, material, and orientation—tasks that typically require specialized sensors.

A. Evaluating Sensor Resolution Across Object Size

In this experiment, we evaluate the spatial resolution of the acoustic sensor by distinguishing objects of slightly different sizes. We 3D-print 16 cubes with edge lengths ranging from 10 mm to 30 mm using Polylactic Acid (PLA), keeping material density and shape constant (Fig. 4). Fig. 5 shows predictions for both known and unknown cubes, with the

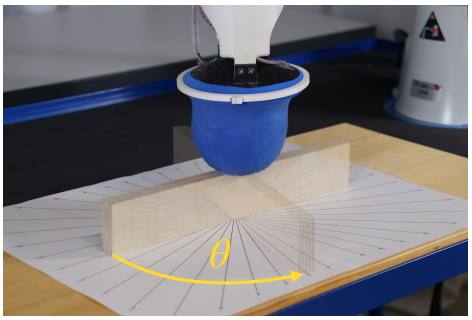

Fig. 6. We predict the orientation of a wooden bar from acoustic data alone. The angle θ parameterizes the orientation. We sample data every $9°$ across $180°$, summing to 19 distinct object rotations. The size of the bar is $270\,\mathrm{mm}\times15\,\mathrm{mm}\times44\,\mathrm{mm}$. The figure overlays two photos of different object orientations. Fig. 7 shows the results of the experiment.

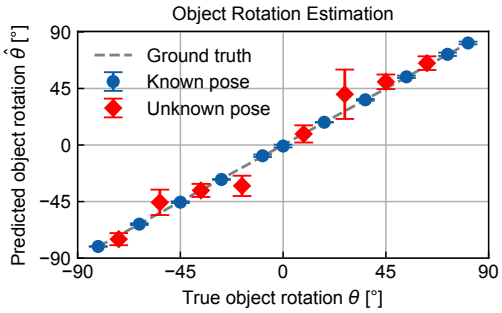

Fig. 7. The acoustic sensor accurately senses an object's orientation. We predict the rotation of a wooden bar inside the gripper as parametrized by an angle θ. The average validation error across known object orientations is $0.6°$ and the test error on orientations not seen at training is $8.0°$.

latter not seen during training. The sensor achieves a small overall error of $2.6\,\mathrm{mm}$. The largest error occurs for the smallest cube, likely because the acoustic modulations are weak as the cube size ($10\,\mathrm{mm}$) approaches the granular medium diameter ($6\,\mathrm{mm}$).

How does this sensor resolution compare to related work? Camera-based tactile sensing for jamming grippers can achieve high, pixel-level spatial resolution [5]. However, approaches relying on flexible sensor arrays attached to the gripper membrane presumably have lower spatial accuracy, on the order of a few centimeters. For example, Mo et al. [9] use a 3×3 tactile array on an $80\,\mathrm{mm}\times80\,\mathrm{mm}$ membrane. They do not explicitly report spatial resolution, but this layout implies an effective minimum resolution of roughly $27\,\mathrm{mm}$. By this measure, our acoustic sensor provides an order-of-magnitude higher spatial resolution while fully preserving gripper compliance and maintaining a simple design.

B. Evaluating Sensor Robustness to External Acoustic Noise

Universal jamming grippers are used in unstructured environments and may be exposed to external acoustic noise. Therefore, we test the generalization of the trained object size predictor to external noise. We place a consumer-grade loudspeaker at approximately $40\,\mathrm{cm}$ from the gripper, play white noise, and measure the noise level at the gripper using

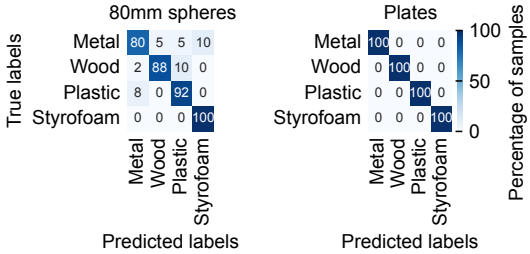

Fig. 8. The acoustic sensor detects the materials of $80\,\mathrm{mm}$ spheres with a high accuracy of $90\,\%$ (left). It perfectly classifies plates of different materials (right). Fig. 4 shows the corresponding objects. We place these objects on a concave rubber piece, preventing the spheres from rolling.

a mobile phone. We then reevaluate the sensor's capability to measure object size. Our results show that the RMSE rises from $2.7\,\mathrm{mm}$ at $45\,\mathrm{dBA}$ ambient noise levels to $3.5\,\mathrm{mm}$ at $80\,\mathrm{dBA}$ external noise. The small error increase of $0.8\,\mathrm{mm}$ under substantial noise is surprising but aligns well with the results by Wall et al. [17], which also showed good robustness against external noise. The membrane seems to act as a good insulator for external noise, making the introduced noise inside the cavity relatively smaller than the reflected speaker signal level.

C. Evaluating Sensor Resolution Across Object Orientation

Sensing the object orientation within the gripper immediately before grasping is challenging for vision-based methods due to self-occlusion. Therefore, we evaluate the acoustic sensor's resolution across another dimension: object orientation. In the experiment in Fig. 6, we record acoustic data of an object in 19 unique rotations and attempt to predict its orientation as represented by an angle θ. As shown in Fig. 7, our method achieves high-resolution orientation estimation with a validation error of only $0.6°$. Notably, we withhold 8 of the 19 object rotations for testing—an unusually large test split of $42\,\%$—yet the model still generalizes well to unseen rotations with a test error of $8.0°$. These results demonstrate that the acoustic sensor can achieve high-resolution orientation estimation, complementing its size and material sensing capabilities.

D. Evaluating Sensor Discrimination of Object Materials

We hypothesized that part of the speaker's acoustic energy propagates *into* the object and reflects back into the gripper cavity toward the microphone. Therefore, the sensor should be able to discriminate *intrinsic* object properties such as material. To test this, we classify $80\,\mathrm{mm}$ spheres and $\sim5\,\mathrm{mm}$ thick plates made of metal, wood, plastic, and Styrofoam (Fig. 4). The results in Fig. 8 show that the spheres are distinguished with $90\,\%$ accuracy, while plates can be perfectly differentiated with $100\,\%$ accuracy.

An additional experiment with smaller, $30\,\mathrm{mm}$ spheres yielded accuracies near random chance. Because material information lies *inside* the object, energy must be transferred into the object and back to the gripper. Smaller spheres have a reduced contact area with the membrane, limiting this energy transfer, possibly explaining the lower classification

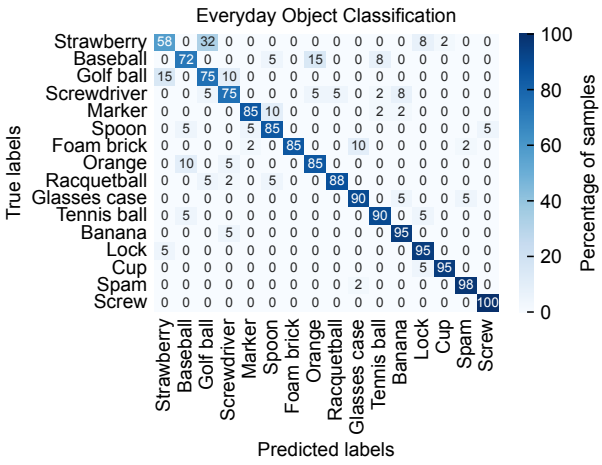

Fig. 9. After showing strong performance on simple shapes, the sensor also performs well on realistic objects: it classifies 16 everyday objects from the YCB [21] dataset with an avg. validation accuracy of 85.6 %. Some objects with similar shape and size are occasionally confused (e.g., strawberry and golf ball, or baseball and orange). Fig. 4 shows the corresponding objects.

performance. In contrast, object size prediction on small objects (10 mm to 30 mm) worked well, suggesting that predicting size relies less on energy exchange between object and gripper. Presumably, object size can be estimated directly from the object's imprint in the granular medium.

In summary, acoustic sensing allows accurate material prediction for sufficiently large objects. This capability is particularly useful in unstructured environments, because cameras cannot differentiate visually identical objects that differ in other physical properties like material.

V. Sensor Application in Realistic Setting

The previous section showed the efficacy of the acoustic jamming gripper in predicting various physical quantities of simple objects like cubes, plates, and spheres. This section applies and discusses the approach in a realistic setting when classifying complex, everyday objects.

A. Application to Object Classification and Sorting

In this section, we demonstrate the system's applicability in a real-world sensing and grasping task. A natural choice is a purely tactile object sorting task, commonly found in industrial settings where vision is impaired due to occlusion. First, we need a more diverse object dataset including objects of different size, shape, and softness. The YCB [21] dataset is suitable because it provides a wide range of objects along exactly these dimensions. However, not all objects in this dataset fit inside the gripper. Therefore, we select some objects that fit inside the gripper and are also graspable. Fig. 4 shows the 16 everyday household objects we selected from the larger YCB dataset. Our object set features deformable objects (foam brick, tennis ball, racquetball), small objects (screw), large objects (spam can, glasses case), and objects of similar shape (golf ball, racquetball, baseball, tennis ball, plastic orange, plastic strawberry).

Fig. 10. Our purely tactile object sorting task shows that while objects are classified correctly from sound, the sensing hardware fully preserves the gripper's compliant grasping. The supplementary video shows the robot reliably sensing the objects' class, grasping, and sorting them into the correct bin. The dashboard on the screen visualizes the object probabilities.

In realistic environments, object pose uncertainty may be large. To make our models robust against object pose changes, we record data in 20 random gripper poses for each object ($\pm$ 90° z-rotation and $\pm$ 1 cm translation). Since audio signal variance between poses is large but samples within the same pose are nearly identical, we record only two one-second samples per pose to avoid overfitting. A dataset for 16 objects consists of $\sim$ 11 minutes of audio (16 objects $\times$ 20 poses $\times$ 2 samples $\times$ 1 s $=$ 640 s), taking roughly one hour to record. Fig. 9 shows that our object classifier achieves 90 % average validation accuracy or more for seven objects and 85.6 % overall. The sensor only confuses some objects of similar geometry and size (e.g., strawberry and golf ball, or baseball and orange).

Finally, we demonstrate the system's strong sensing and grasping performance in a real-world tactile object sorting task. Fig. 10 shows a still from the supplementary video. The video highlights that the acoustic sensor fully preserves the gripper's compliance and grasping capabilities while objects are classified correctly purely based on sound. We achieve 53 minutes of uninterrupted sorting of four objects: strawberry, marker, lock, and screw. The operator places the objects in a randomized order at known grasping locations. The gripper then approaches, classifies, and sorts each object into its corresponding bin. We complete 39 successful sorts in a row, each requiring a correct prediction, a stable grasp, and a controlled release. The demonstration ends at the 40th classification because no object probability crosses the user-defined minimum confidence threshold of 60 % as lock and strawberry receive similar class probabilities. Nevertheless, this last prediction would have also been correct. The video also demonstrates robustness to slight variations in pose as the operator sets down the objects in poses that are likely not observed during training. The gripper always manages to secure a stable grasp and ensures no object drops. In summary, our results imply that the acoustic sensor can support differentiating objects in realistic, cluttered environments such as warehouses, where reliance on vision is not always possible due to occlusions or insufficient lighting.

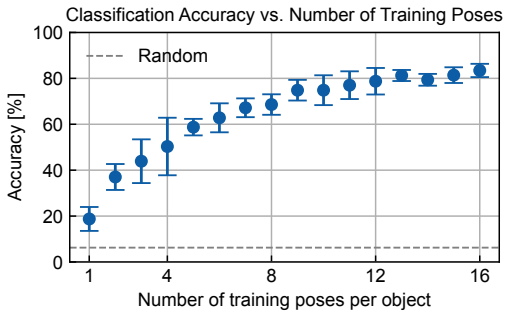

Fig. 11. Reaching adequate YCB [21] object classification accuracies of roughly 80 % requires sampling training data in at least 12 different poses per object. Our best model trains on data from 16 different poses.

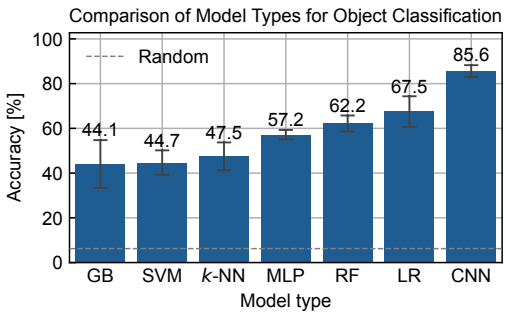

Fig. 12. Classifier performance on raw STFT input (*GB*: gradient boosting, *SVM*: support vector machine with RBF kernel, *k-NN*: k-nearest neighbors, *MLP*: multi-layer perceptron, *RF*: random forests, *LR*: multi-class logistic regression, *CNN*: our best-performing convolutional neural network)

B. Data Requirements and Model Baselines

The practical deployment of any learning-based robotic sensing system is constrained by the cost of data collection. To better understand the trade-offs in our acoustic sensing approach, we analyze data requirements and compare model architectures using the largest and most diverse dataset recorded in this work: the 16-object YCB set.

We first investigate how much training data is required to achieve good classification performance. Our dataset contains 20 poses per object (16 training, 4 validation poses), with two samples per pose. In Fig. 11, we evaluate how classification accuracy changes as we vary the number of unique poses per object in the training set. The results show that we need 12 gripper poses to reach adequate classification accuracies of approximately 80 %. While the performance improves sharply when increasing from very few poses, adding more poses continues to yield gains but with diminishing returns. This analysis provides practical guidance for the effort required to collect training data.

Next, we evaluate different classifiers on the same STFT feature vectors to understand the impact of model choice. Fig. 12 shows that our CNN achieves the highest accuracy of 85.6 %, averaged over all cross-validation folds, clearly outperforming other baselines. Presumably, the lower performance of the other classifiers is partly due to the high dimensionality of the input features (1025 dimensions). In such high-dimensional input spaces, distance metrics lose their meaning due to the curse of dimensionality. Therefore, methods based on computing distances (such as k-NN and SVM [22]) suffer from the reduced discriminative power of distance metrics. While tree-based and linear models (RF, GB, LR) perform better than random chance, the CNN still clearly outperforms them.

C. Limitations

We now discuss the limitations of the acoustic sensing approach. Acoustic data is high-dimensional and complex, making analytical modeling of these phenomena challenging. We therefore employ machine learning to uncover patterns in the data. This is the strength of the approach, but also its weakness, as it inherits the limitations of machine learning.

For instance, the approach is data-dependent as the models presented so far are supervised. Hence, new training data is required for new object classes or modalities we wish to predict, but collecting data in the real world is costly. A mitigation would be to learn more general object representations—such as object shape—to directly analyze novel objects without task-specific datasets. The success of vision and language models was mainly due to the availability of vast amounts of diverse training data. However, the presented acoustic sensing approach is highly morphology-specific, which may limit scalability. For example, collecting data cannot be easily parallelized on multiple grippers, because cross-gripper transfer remains challenging.

Like all machine learning models, our approach is susceptible to shortcut learning, where the model exploits spurious correlations in the training data rather than the true causal features. For instance, in our object orientation prediction experiments in Section IV-C, we initially rotated the gripper to automate data collection instead of rotating the object. This caused the model to overfit to the sound of the arm's motors and fail to generalize to actual object rotations. By rotating the object during training data collection, we achieved good generalization, highlighting the importance of careful dataset collection to avoid unintended shortcuts.

Moreover, our models perform well on a given gripper; however, model transfer to *different* gripper morphologies remains challenging. For instance, replacing or repositioning the latex balloon alters the acoustic properties of the cavity, which can reduce the reliability of trained models. The core difficulty is that *any* morphological change introduces an acoustic distribution shift: when caused by object contact, such shifts carry useful tactile information, but when caused by gripper design changes, they instead introduce an undesired acoustic shift. Because the models cannot separate these cases, they do not transfer well across gripper designs. These results align with the work by Wall et al. [17], who report that models transfer poorly between different soft robot fingers.

We do not view sensitivity to changes in gripper morphology as a fundamental challenge but as an *entanglement* problem. The reflected sound jointly encodes multiple factors: object state (size, material, orientation), gripper

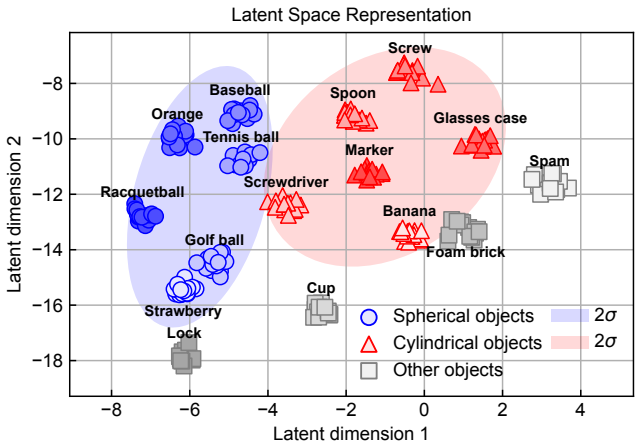

Fig. 13. Disentangled representations can be learned from acoustic data without class labels. The learned latent space separates object properties from pose: objects with similar shape cluster together regardless of pose—e.g., *spherical* objects (orange, baseball, tennis ball, racquetball, golf ball, strawberry) and *cylindrical* objects (screw, spoon, glasses case, marker, screwdriver, banana). The shaded ellipses denote Gaussian covariance contours at $2\,\sigma$. Each data point corresponds to a sample in a different object pose. This plot shows training data points for better readability; Figs. 14 and 15 provide quantitative results on validation data.

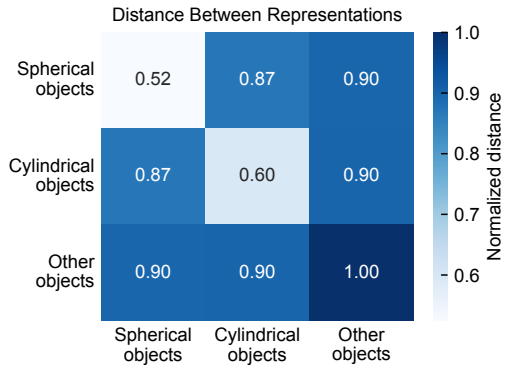

Fig. 14. Our quantitative analysis confirms the semantic structure in the latent space: intra-category distances for *spherical* and *cylindrical* objects are smaller than inter-category distances, showing that the representations cluster objects by shape. *Other* objects span diverse geometries and therefore remain well separated, as reflected in their large intra-category distance. While Fig. 13 is a snapshot of one fold, we average over all folds of the k-fold cross-validation here, providing a more comprehensive view across the entire dataset and different initializations.

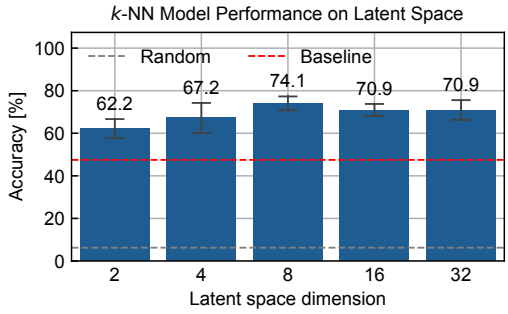

Fig. 15. Disentangled embeddings improve object separability: applying k-NN on the latent space yields up to 26.6 % higher accuracy than on raw STFT features, showing that the learned representations expose relevant variations while ignoring irrelevant factors like pose. The k-NN baseline is 47.5 % from Fig. 12. Performance peaks at eight latent dimensions, likely reflecting a balance between capturing all meaningful variations in the acoustic signal while avoiding overfitting to irrelevant factors.

morphology (membrane placement, filling, jamming state), and environmental conditions (ambient noise, contact force, pump vibrations). So far, we have shown that variations in *one* of these factors—the object state—can be reconstructed via morphological sensing. In principle, with appropriate data and inductive biases, machine learning should be able to disentangle *all* of these generative factors. This perspective motivates the search for disentangled acoustic representations, which we explore in the next section.

D. Disentangled Representations for Morphological Sensing

The acoustic signal reflects multiple entangled factors, including object state, gripper morphology, and environmental conditions. Disentangled representations aim to factorize these variations, aligning each axis in a latent space with a semantically meaningful factor. Invariant subspaces—such as pose-invariant or gripper-invariant subspaces—can be understood as slices through this higher-dimensional disentangled space, where irrelevant axes are marginalized. Learning such lower-dimensional representations is desirable because they isolate task-relevant variations (e.g., object properties) while ignoring irrelevant ones (e.g., object pose or gripper design), enabling robustness to such irrelevant factors.

We show how such representations can be discovered without semantic labels by exemplifying how a pose-invariant subspace can be found. Using self-supervised dimensionality reduction [23], we map measurements of the same object in different poses close together in the latent space, while separating measurements of different objects. This approach requires no object class labels, highlighting a practical advantage: disentangled representations can be learned directly from the data without expensive labeling.

Our experiments show that the latent space in Fig. 13 captures pose-invariant object properties: each object occupies a distinct region regardless of pose and a semantic structure emerges. *Spherical* and *cylindrical* objects gather in common regions, and spherical objects of similar size cluster together (e.g., orange, baseball, and tennis ball group together, as do golf ball and strawberry), while *other* objects with diverse properties cover different parts of the space. The analysis in Fig. 14 confirms this semantic structure: intra-category distances for *spherical* and *cylindrical* objects are smaller than inter-category distances. Objects with diverse geometries, such as those in the *other* category, are correctly kept apart in the latent space, reflected in their large intra-category distance. Finally, Fig. 15 shows the latent space improves separability of the acoustic data: a simple k-NN classifier on the disentangled embeddings achieves up to 26.6 % higher accuracy than on raw STFT features, confirming that the latent space exposes meaningful semantic structure while ignoring task-irrelevant variations, such as object pose.

In principle, the same approach can be extended to discover other invariant subspaces. For example, a gripper-invariant subspace can be obtained by ensuring that variations in gripper morphology (e.g., membrane placement, filling, etc.) do not change the embedding for the same object. Achieving this requires collecting repeated measurements across different gripper configurations. Overall, these findings demonstrate that disentangled subspaces exist within the high-dimensional acoustic data and that they can be discovered from data alone. By appropriately constraining the learning process, one can extract subspaces that are invariant to specific nuisance factors, enabling more robust and transferable acoustic sensing.

VI. Conclusion

We demonstrated that the soft body of a universal jamming gripper, typically a challenge for sensorization, can be turned into an asset through morphological sensing. In this framework, the gripper's morphology acts as a resonant body, capturing object properties as acoustic variations.

We made three key contributions. First, we showed that acoustic sensing is a suitable morphological sensing approach for jamming grippers, as the rigid hardware can be placed away from the deformable parts, fully preserving the gripper's compliance. Our realistic object sorting task confirmed that grasping performance is maintained, achieving 53 minutes of purely tactile sorting without dropping a single object. Second, we demonstrated that a single sensor can accurately reconstruct various object properties, including size, material, orientation, and class, with predictions robust to external noise due to the shielding effect of the gripper membrane. Third, we showed that disentangled representation spaces exist within the high-dimensional acoustic data. These latent spaces separate task-relevant information, such as object properties, from irrelevant variations, like object pose, enabling more robust predictions. Overall, our work calls for a perspective shift: a robot's body is not merely a passive structure, but an active participant in perception, realized through morphological sensing.

References

[1] L. Weber, T. Wienert, M. Splettstößer, A. Koenig, and O. Brock, "Acoustic sensing for universal jamming grippers," in *IEEE International Conference on Robotics and Automation (ICRA) Workshop on Acoustic Sensing and Representations for Robotics*, 2025.

[2] E. Brown, N. Rodenberg, J. Amend, A. Mozeika, E. Steltz, M. R. Zakin, H. Lipson, and H. M. Jaeger, "Universal robotic gripper based on the jamming of granular material," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 107, no. 44, pp. 18 809–18 814, 2010.

[3] V. Wall, "Morphological sensing for soft pneumatic actuators based on acoustics and strain," Dissertation, Technische Universität Berlin, 2024.

[4] X. Meng, H. Xi, J. Han, and A. Song, "A universal jamming gripper for visual-based grasping in narrow spaces," in *IEEE International Conference on Mechatronics and Automation (ICMA)*, 2023, pp. 2360–2365.

[5] S. Li, X. Yin, C. Xia, L. Ye, X. Wang, and B. Liang, "TaTa: A universal jamming gripper with high-quality tactile perception and its application to underwater manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2022, pp. 6151–6157.

[6] T. Sakuma, F. Von Drigalski, M. Ding, J. Takamatsu, and T. Ogasawara, "A universal gripper using optical sensing to acquire tactile information and membrane deformation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1–9.

[7] T. Sakuma, T. Kiyokawa, T. Matsubara, J. Takamatsu, T. Wada, and T. Ogasawara, "Jamming gripper-inspired soft jig for perceptive parts fixing," *IEEE Access*, vol. 11, pp. 62 187–62 199, 2023.

[8] L. Y. W. Loh, U. Gupta, Y. Wang, C. C. Foo, J. Zhu, and W. F. Lu, "3D printed metamaterial capacitive sensing array for universal jamming gripper and human joint wearables," *Advanced Engineering Materials*, vol. 23, no. 5, p. 2001082, 2021.

[9] L. Mo, W. Xie, J. Qu, J. Xia, Y. Li, Y. Zhang, T. Ren, Y. Yang, J. Yi, C. Wu, and Y. Chen, "Empowering particle jamming soft gripper with tactility via stretchable optoelectronic sensing skin," *Advanced Intelligent Systems*, vol. 7, no. 1, p. 2400285, 2024.

[10] S. Y. Liang and D. A. Dornfeld, "Tool wear detection using time series analysis of acoustic emission," *Journal of Engineering for Industry*, vol. 111, no. 3, pp. 199–205, 1989.

[11] J. Lee, H. Choi, D. Park, Y. Chung, H.-Y. Kim, and S. Yoon, "Fault detection and diagnosis of railway point machines by sound analysis," *Sensors*, vol. 16, no. 4, p. 549, 2016.

[12] A. Glowacz, "Acoustic fault analysis of three commutator motors," *Mechanical Systems and Signal Processing*, vol. 133, p. 106226, 2019.

[13] J. Sinapov, M. Wiemer, and A. Stoytchev, "Interactive learning of the acoustic properties of household objects," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2009, pp. 2518–2524.

[14] K. Xu and J. Chan, "Soft tactile sensors for robot grippers using acoustic sensing," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025, pp. 10 423–10 430.

[15] S. Lu and H. Culbertson, "Active acoustic sensing for robot manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 3161–3168.

[16] K. Zhang, D.-G. Kim, E. T. Chang, H.-H. Liang, Z. He, K. Lampo, P. Wu, I. Kymissis, and M. Ciocarlie, "VibeCheck: Using active acoustic tactile sensing for contact-rich manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025, pp. 12 278–12 285.

[17] V. Wall, G. Zöller, and O. Brock, "Passive and active acoustic sensing for soft pneumatic actuators," *The International Journal of Robotics Research (IJRR)*, vol. 42, no. 3, pp. 108–122, 2023.

[18] U. Yoo, Z. Lopez, J. Ichnowski, and J. Oh, "POE: Acoustic soft robotic proprioception for omnidirectional end-effectors," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 14 980–14 987.

[19] M. S. Sofla, H. Golshanian, V. Rajendran S, and A. Ghalamzan E, "Soft acoustic curvature sensor: Design and development," *IEEE Robotics and Automation Letters (RA-L)*, vol. 9, no. 11, pp. 9518–9525, 2024.

[20] C.-W. Kang, K. Hashitsume, and E.-S. Jang, "Investigation of the sound-absorbing performances of pure coffee grounds," *BioResources*, vol. 18, no. 2, pp. 3308–3318, 2023.

[21] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The YCB object and model set: Towards common benchmarks for manipulation research," in *IEEE International Conference on Advanced Robotics (ICAR)*, 2015, pp. 510–517.

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[23] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2006, pp. 1735–1742.