

Supplementary Material: SonicSense: Object Perception from In-Hand Acoustic Vibration

Anonymous Author(s)

Affiliation

Address

email

1 A. Hardware Configuration of Acoustic Robot Hand

2 Our hardware is built by 3D printing with Polylactic Acid (PLA) filament. We use the LX-224
3 servo motor to actuate the finger which provides 20kg-cm torque and accurate position and voltage
4 feedback. The contact microphones we used are commercially available on Amazon and the two
5 counter weights on each fingertip is 40g. We place the controller of the motors as well as the audio
6 jack inside the palm. The cost split of building our acoustic robot hand is shown in Tab. 1.

Part	Amount	Price in total (\$)
Lx224 motor	4	75.96
TTL/USB Debugging Board	1	12.99
4pcs piezo contact microphone	1	28.58
Audio cable	4	59.96
Counterweight	8	27.44
PLA 3D printing material (689.11g)	1	10.33
Total:		215.26

Table 1: Cost of the acoustic robot hand.

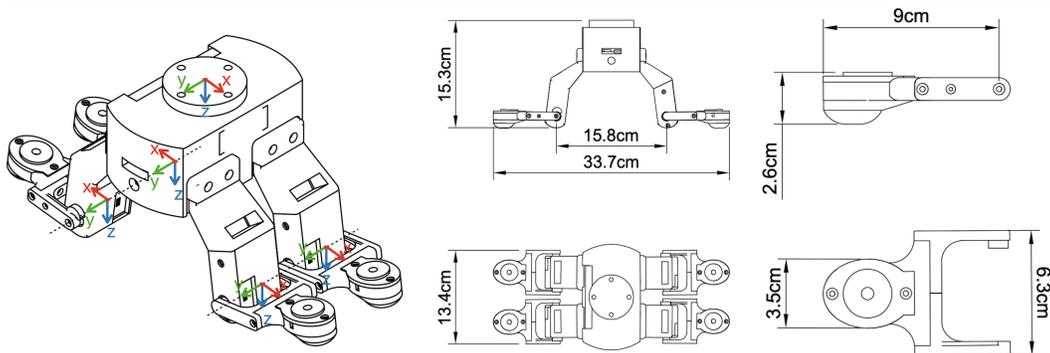


Figure 1: CAD model and coordinate system of the robot hand.

7 Contact Position Calculation

8 We will introduce the calculation of the approximated contact point location in this section. We
9 set the fingertip position at the center of the fingerprint. Once the binary contact event is detected,
10 we will record the angle of each finger joint. Based on our CAD model, we can also obtain other
11 kinematic parameters of the hand such as the length of each link. Fig. 1 shows the sketch of our
12 hardware design and its coordinate system. We use the center of the palm as the origin of the hand.
13 The y-axis is parallel to the four finger joints and the z-axis is perpendicular to the palm, facing
14 towards the finger. The coordinate system of the finger joints has the same orientation as the above
15 central coordinate system of the hand. The center of the finger joint coordinate is located at the

16 center of each motor. We denote the position of joint 1 as $p_1 = \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix}$, where $x_1 = 7.845cm$,
 17 $y_1 = 3.429cm$, $z_1 = 13.691cm$. Considering that the four fingers are symmetrical, the positions of
 18 the rest of the fingers can be represented as

$$p_2 = \begin{bmatrix} x_1 \\ -y_1 \\ z_1 \end{bmatrix}, p_3 = \begin{bmatrix} -x_1 \\ y_1 \\ z_1 \end{bmatrix}, p_4 = \begin{bmatrix} -x_1 \\ -y_1 \\ z_1 \end{bmatrix} \quad (1)$$

19 We use Pe_i to represent the fingertip position under its joint coordinate. Since the length between
 20 the origin of joint coordinate and the fingertip is $l = 7.6cm$, the fingertip position of finger i under
 21 the robot hand coordinate, denoted as ${}^H Pe_i$, can be calculated by

$${}^H Pe_i = p_i + RPe_i \quad (2)$$

22 Here, $Pe_i = \begin{bmatrix} l \\ 0 \\ 0 \end{bmatrix}$. For finger 1 and finger 2, R can be represented as:

$$R = R_y(-\theta_i + 4.5^\circ) \quad (3)$$

23 The θ_i here represents the joint angle of finger i . For finger 3 and finger 4, R can be represented as:

$$R = R_y(\theta_i - 4.5^\circ)R_z(180^\circ) \quad (4)$$

24 We attach the robot palm on the end effector of the Franka Emika Panda arm and align the end effec-
 25 tor coordinate with the robot hand coordinate to ensure ${}^H Pe_i = {}^E Pe_i$. Having the transformation
 26 matrix of the end effector coordinate to the robot arm ${}^R E T$, we can then calculate the contact position
 27 under the robot arm coordinate, represented as ${}^R Pe_i$, by:

$${}^R Pe_i = {}^R E T {}^E Pe_i \quad (5)$$

28 B. Interaction Policy for Data Collection

29 We will introduce the interaction policy for the robot to conduct real-world data collection in this
 30 section. Since vision is not used, the dimension of the object is unknown. Therefore, the first
 31 challenge of the interaction policy is to estimate the dimension of the object.

32 We fix the object at the center of a black base on top of a wooden board as shown in Fig. 2A. The
 33 robot will first perform two tapping motions from the side with two different heights (Fig. 2B and
 34 2C) and save the valid contact points. The tapping motion is able to make contact with objects since
 35 the objects' body is across the centerline of the black base. Based on the highest detected contact
 36 point location, the robot will have a rough estimation of the height of the object.



Figure 2: **Initial estimation stage of the interaction policy.** (A)The black base and its centerline. (B)The first side tapping in a lower position. (C)The second side tapping in a higher position.

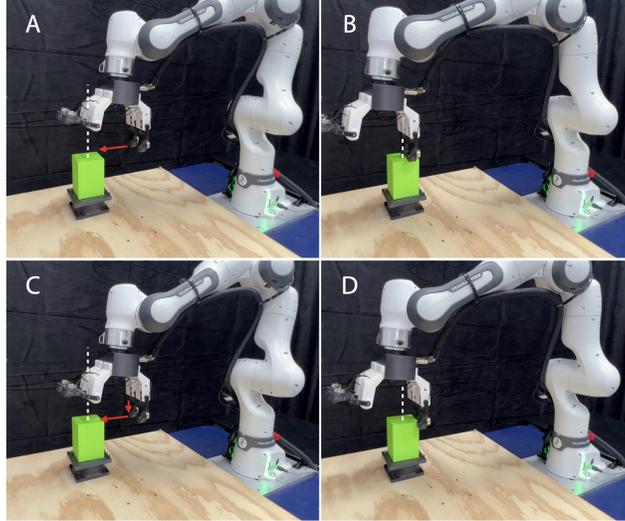


Figure 3: **Exploration of highest edge.** (A)The third finger gets to a position higher than the roughly estimated height and moves toward the centerline of the black base. (B)The third finger reaches the centerline and no contact happens. (C) The third finger goes back to the previous initial position as shown in (A), moves down a little bit, and moves toward the centerline again. (D)The third finger makes contact with the object. A contact event is detected and the contact position is saved.

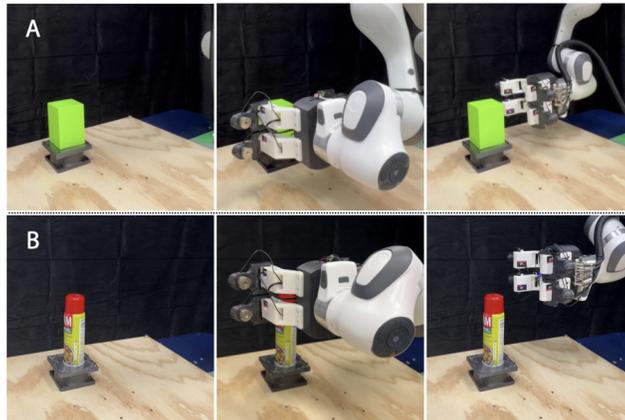


Figure 4: **Two scales of left side and back side tapping.** (A) The lower starting position of these two tapping motions on a smaller object. (B)The higher starting position of these two tapping motions on a larger object.

37 To obtain a more accurate estimation of height and radius of the object, the robot will reset to its
 38 initial home position to lift its two fingers in front, and use its third finger to explore the highest
 39 edge on top of the object. Specifically, starting from a suitable height away from the objects, the
 40 third finger of the hand moves towards the centerline of the black base (Fig. 3A) and reaches another
 41 side of the centerline (Fig. 3B). The robot keeps monitoring the acoustic signal from the third finger
 42 during this process. Once a contact event is detected, the robot will record the contact position
 43 through the kinematic model and stop exploring. If the third finger does not make contact with the
 44 object and there is no contact event being detected, the robot hand will move back to the initial
 45 position, and move down a little bit as shown in Fig. 3C and keep executing the same edge exploring
 46 motion until a contact sound is detected as shown in Fig. 3D. We use the distance between the
 47 contact point and the centerline as the estimated radius of the object and the distance between the
 48 contact point and the black base plane as the final estimated height of the object.

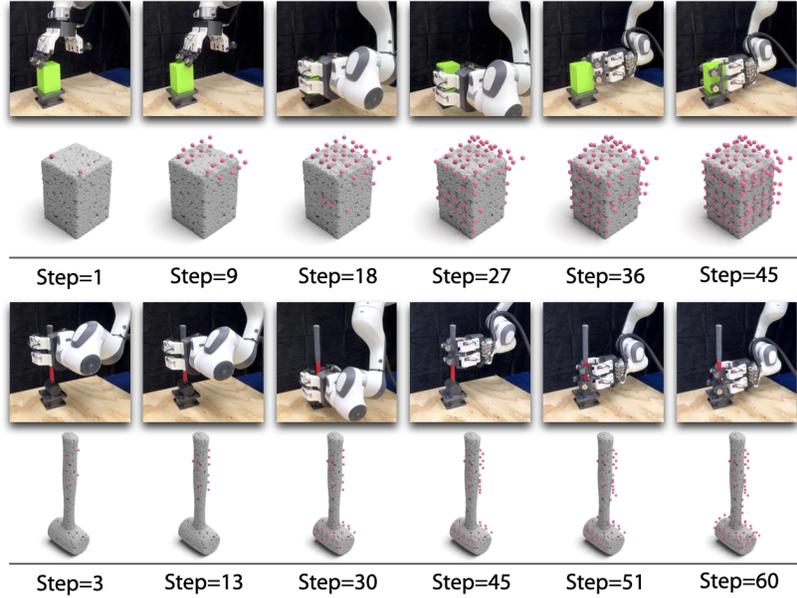


Figure 5: **Visualizations of contact events through tapping motion.** We show the sequence of tapping interactions on a smaller object in the first row and a larger object in the second row. The x-axis shows the step progression. The estimated contact fingertip positions are shown as points overlaid on top of the ground-truth 3D shapes.

49 Following the estimation, the robot will interact with the object through tapping motions on three
50 sides of the object in sequences including the top side, the left side, and the backside of the object.
51 During each tapping sequence, the robot hand will approach and tap the object step by step, ensuring
52 continuous tapping while gradually transitioning from top to bottom (or left to right for top tapping)
53 to cover the entire surface of the object. A detailed illustration of the tapping motion is provided in
54 Fig. 5. The robot relies on the estimated radius to determine whether to collect tapping data on top
55 of the object since the useful information is limited when the surface area on top of the object is too
56 small. We set the radius threshold to be 2cm in this case. Additionally, depending on the estimated
57 height of the objects, the robot decides to start from a higher position or a lower position for two
58 different scales of data collection during the left-side tapping and back-side tapping as illustrated in
59 Fig. 4. We set the height threshold to be 20cm. For a more comprehensive visual demonstration,
60 please refer to our Supplementary Movies. When our robot hand taps objects, we can determine
61 such contact events from the voltage feedback from the motor in each finger. We put a 10Ω resistor
62 in series with the servo motor power. When the motor encounters external force, it receives an
63 increase of current supply and, as a result, a decrease of voltage from the motor voltage sensor
64 readings. Though we also experimented with detecting salient acoustic signals to detect the binary
65 contact event, we found that the natural motor voltage feedback embedded by the motors provides
66 more reliable and accurate signals. This is because the voltage feedback can reflect subtle resistance
67 encountered by the robot finger, where acoustic vibrations from the contact microphone also include
68 motor noises. After the contact event is detected, we can localize the contact points in 3D space with
69 respect to the robot with the forward kinematics of our robot hand and arm.

70 We conducted this real-world data collection on the 83 real-world objects mentioned in our main
71 texts and used the tapping data only with valid contact events for our material classification, shape
72 reconstruction, and object re-recognition tasks. The tapping data are saved in txt file format, includ-
73 ing position in the x-axis (x), position in the y-axis (y), position in the z-axis (z), binary tapping
74 detection from acoustic vibration signal (a), binary tapping detection from voltage signal (v), finger
75 index (f), and index of tapping (i):

$$\{x, y, z, a, v, f, i\} \quad (6)$$

76 Each tapping data is also accompanied by a five-second audio recording, which is guaranteed to
77 cover the acoustic signals of the impact sound. We will present the audio processing procedure in
78 the next section.

79 **C. Acoustic Vibration Recording and Processing**

80 To obtain more useful information, we first extracted the striking vibration signal from the 5-second
81 recording. Typically, the anticipated tapping vibration signal exhibits a noticeable characteristic,
82 often resembling a peak-shaped pattern. However, real-world acoustic vibration signals may contain
83 random noises, so simply extracting the signal around maximum amplitude will not work. To extract
84 informative acoustic vibration signals, we used a window of 1000 units of acoustic waveform to
85 monitor the average absolute amplitude of the acoustic signals. If the average amplitude in the
86 current window is larger than both the average amplitude in the previous window and the next
87 window, 20000 units of the acoustic signals ($20000/44100Hz = 0.45351474s = 453.5147ms$)
88 will be extracted around the beginning timestamp of the current window. We applied the same
89 extracting mechanism to the recording with no obvious striking signal, for example, foam objects.
90 For those objects, most extracted signals are motor noises. However, we consider this as a feature
91 of the soft materials perceived by our robot hand. Finally, we followed previous work to convert the
92 raw audio signals to Mel spectrogram representations with a dimension of 64×64 resolution. The
93 length of the Fast Fourier transform (FFT) window is set to 2048 and the number of the Mel bands
94 is set to 64. The highest frequency is restricted to 8192Hz. To ensure the correct dimension of the
95 Mel spectrogram along the temporal dimension, the number of samples between consecutive frames
96 is rounded to the nearest value obtained by dividing the total number of audio samples by 64.

97 **D. Implementation Details for Material Classification**

98 **Dataset Construction**

99 In this section, we explain how we obtained the material label for each of the contact positions
100 around the object. During tapping data collection, we used two depth cameras to capture two point
101 clouds of the object after fixing the object on the platform. We fuse those two point clouds and use
102 the annotated point cloud model to align with the object. We then searched the nearest point of the
103 annotated point cloud for each of the contact points and assigned the annotated material label to that
104 contact point based on the label of the nearest point. The material label serves as the ground truth
105 for training our material classification model.

106 Tapping data from 82 objects is used for the material classification task. We removed the acoustic
107 data collected from a disinfecting wipe because the paper material is too soft and there is a thin
108 plastic film attached to the paper, so the signal is too unique and out of distribution from all other
109 paper materials. We conducted separate training on three different random splits of the dataset to
110 evaluate our method. Note that our splits are based on objects to avoid strong overlapping between
111 training, validation, and testing data. We have 60 objects for training, 11 objects for validation, and
112 11 objects for testing. We balanced the number of acoustic data in each of the material classes for
113 training and validation by duplicating them from random choice. Therefore, due to different objects
114 coming with different sizes and different numbers of tapping data, the number of data points for
115 each split will be different under the same number of objects. The resulting number of data points
116 in each split is shown in Tab. 2.

117 **Material Classification Model Architecture**

118 We show the detailed architecture design of our material classification network in Tab. 3.

Dataset splits	Training	Validation	Testing
Split 1	8829	1521	948
Split 2	7767	2142	1293
Split 3	8433	1278	1195

Table 2: **Material classification data statistics.**

Layer name	Input Channel	Output Channel	Kernal Size	Stride Size	Padding
Conv1	1	16	6	2	0
Batch norm + Relu	N/A	N/A	N/A	N/A	N/A
Maxpool1	16	16	2	2	0
Conv2	16	32	5	1	0
Batch norm + Dropout+ Relu	N/A	N/A	N/A	N/A	N/A
Maxpool2	32	32	2	2	0
Conv3	32	150	5	1	0
Batch norm + relu+dropout	N/A	N/A	N/A	N/A	N/A
Fc1	150	70	N/A	N/A	N/A
Dropout	N/A	N/A	N/A	N/A	N/A
Fc2	70	9	N/A	N/A	N/A

Table 3: **Neural network architecture of the material classification model.**

119 Material Classification Hyperparameters and Results

120 We list all the hyperparameter selections and experimental results across different random data splits
 121 in Tab. 4. We select the model based on the best validation accuracy. Due to the testing dataset being
 122 unbalanced, we use the average F1 score as our evaluation metric. The confusion matrix of the first
 123 experiment is shown in Fig. 6.

	Split 1	Split 2	Split 3
Max epoch	300	300	300
Learning rate	0.00903	0.00935	0.00903
Dropout	0.52105	0.40947	0.30927
Batch size	36	32	32
Optimizer		SGD	
LR schedule		Decays the Learning rate by 0.1 every 200 steps	
Best train accuracy	0.9891	0.9876	0.9826
Best validation accuracy	0.6141	0.5691	0.5696
Refinement parameters (M,K,N)	(8,3,25)	(8,8,25)	(6,1,30)
Testing F1 score	0.5711	0.5035	0.4939
Testing F1 score after refinement	0.8786	0.6924	0.718

Table 4: **Material classification hyperparameters and results on three random data splits.**

124 E. Implementation Details for Shape Reconstruction

125 Dataset Construction

126 We leverage the contact positions collected from the 83 objects for our shape reconstruction task.
 127 The tapping interaction results in highly sparse contact points. The maximum number of contact
 128 points is 300 for each object. We divided the dataset based on 15 shape categories(i.e., bottle, can
 129 cup, hammer, mug, wine glass, cube, cube(concave), cylinder, cylinder(concave), cone, quadrangu-
 130 lar pyramid, triangular pyramid, prism, irregular) and split the dataset for training, validation, and
 131 testing according to these categories. To balance the training dataset for learning, we randomly du-
 132 plicated the data up to 100 for each category. We then randomly sampled 80% to 90% points and
 133 augmented the dataset to 1500 for each shape to augment the training dataset.

134 To further boost the training dataset, we leveraged the PyBullet simulator to collect synthetic tapping
 135 points. We collected 629 synthetic 3D models under the same categories of real-world objects. In

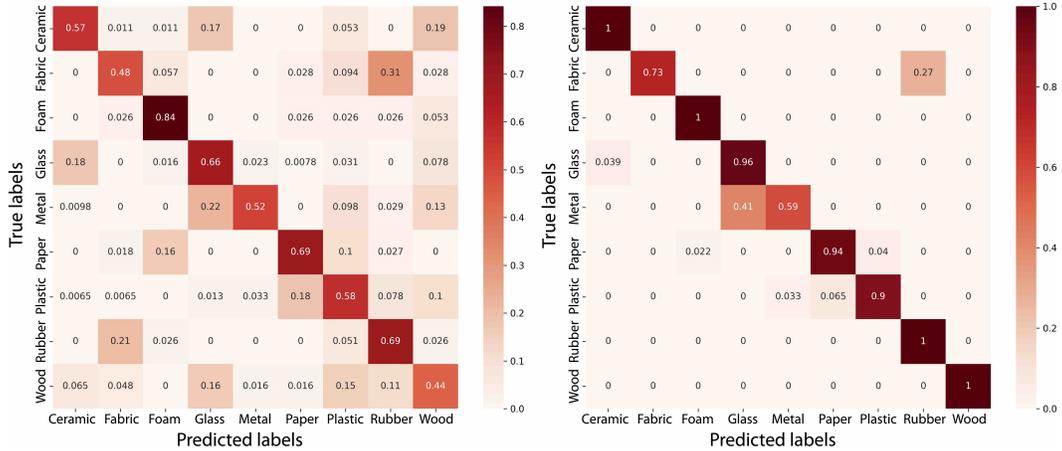


Figure 6: **Confusion matrix of material classification task on one testing dataset** The results of the initial prediction(left) and the results of the prediction after refinement(right).

136 the simulation, We implemented the same interaction policy and data augmentation techniques as in
 137 our real-world data construction. The ground truth model consists of a point cloud model comprising
 138 5000 points for both the real-world objects and the synthetic objects. Our experiments include three
 139 different splits of data with different random seeds where both the validation objects and testing
 140 objects only include real-world objects. For real-world objects, we split them into 61, 11, and 11
 141 objects for training, validation, and testing. In order to leverage the simulation data to improve our
 142 model on our real-world objects, initially, we use all the synthetic datasets to pre-train our model.
 143 Gradually, we blended in more real-world data and reduced the percentage of synthetic data in our
 144 training set. We show the blending schedule in Tab. 5.

Epoch	Synthetic dataset	Real dataset
0-100	100%	0%
100-200	90%	10%
200-300	80%	20%
300-400	60%	40%
400-500	40%	60%
500-600	20%	80%
600-700	10%	90%
700-800	5%	95%
800-1000	0%	100%

Table 5: **The blending schedule of synthetic dataset and real dataset during shape reconstruction training.**

145 **Shape Reconstruction Model Architecture**

146 We show the detailed architecture design of our shape reconstruction network in Tab. 6.

147 **Shape Reconstruction Hyperparameters and Results**

148 We list all the hyperparameter selections and experimental results across different random data splits
 149 in Tab. 7.

150 **F. Implementation Details for Object Re-recognition**

151 **Dataset Construction**

152 We utilize both the tapping vibration signal and contact points from all 82 objects for our object re-
 153 cognition task. For each of the objects, we first randomly split 20% of the tapping data for testing,

Layer name	Input Channel	Output Channel	Kernal Size	Stride Size	Padding
Conv1d	3	128	1	1	0
Batch norm + Relu	N/A	N/A	N/A	N/A	N/A
Conv1d	128	256	1	1	0
Conv1d	512	512	1	1	0
Batch norm + Relu	N/A	N/A	N/A	N/A	N/A
Conv1d	512	1024	1	1	0
Fc1	1024	1024	N/A	N/A	N/A
Relu	N/A	N/A	N/A	N/A	N/A
Fc2	1024	1024	N/A	N/A	N/A
Relu	N/A	N/A	N/A	N/A	N/A
Fc3	1024	3×2000	N/A	N/A	N/A

Table 6: **Neural network architecture of the shape reconstruction model.**

	Split 1	Split 2	Split 3
Max epoch		1000	
Learning rate		0.000005	
Batch size		500	
Optimizer		Adam	
LR schedule		Decays the Learning rate by 0.7 every 500 steps	
Lowest validation loss(CD-L2)	3.6634e-05	5.6421e-05	5.4612e-05
Lowest testing loss(CD-L2)	6.7536e-05	5.6938e-05	6.0342e-05
Lowest testing loss(CD-L1)	0.009184	0.0085316	0.0085615

Table 7: **Shape reconstruction hyperparameters and results on three random data splits.**

154 20% of the tapping data for validation, and 60 % of the tapping data for training. To augment the
155 dataset, 15 tapping data are randomly sampled 500 times for each object in the training dataset, and
156 50 times for each object in the validation and testing dataset. Therefore, there are 410,000 data
157 points for training, 4,100 data points for validation, and 4,100 data points for testing. As in the
158 previous two tasks, we have three different random splits of the dataset for evaluation.

159 Object Re-recognition Model Architecture

160 We show the detailed architecture design of our object re-recognition network in Tab. 8.

161 Object Re-recognition Hyperparameters and Results

162 We list all the hyperparameter selections and experimental results across different random data splits
163 in Tab. 9.

164 G. Implementation Details for Resistance Test Against Ambient Noise

165 We placed a Saramonic LavMicro U1A air microphone next to the fourth finger. We utilized the Na-
166 tional Institute for Occupational Safety and Health (NIOSH) Sound Level Meter App on an iPhone
167 to monitor the noise level. This App has been shown to effectively capture accurate noise levels [1].
168 We stretched the finger so that we can place the iPhone microphone, the fingertip surface, and the
169 air microphone in the same plane. We then set up a speaker facing toward the middle of the robot
170 hand to play Gaussian white noise.

171 H. Parameters of Acoustic Vibration Signal Processing in Characterization 172 Experiments

173 We list all the parameters used to process the acoustic vibration signals for our sensing characteri-
174 zation experiments in Tab. 10. The parameters are used to convert the waveform representation into
175 spectrogram representation as well as the twelve descriptor extraction methods. For liquid object

Layer name	Input Channel	Output Channel	Kernal Size	Stride	Padding
Audio encoder:					
Conv1	15	16	6	2	0
Batch norm + Relu	N/A	N/A	N/A	N/A	N/A
Maxpool1	16	16	2	2	0
Conv2	16	32	5	1	0
Batch norm + Dropout+ Relu	N/A	N/A	N/A	N/A	N/A
Maxpool2	32	32	2	2	0
Conv3	32	150	5	1	0
Batch norm + relu	N/A	N/A	N/A	N/A	N/A
Contact point encoder:					
Conv1d	3	64	1	1	0
Batch norm + Relu	N/A	N/A	N/A	N/A	N/A
Conv1d	64	64	1	1	0
Conv1d	128	128	1	1	0
Batch norm + Relu	N/A	N/A	N/A	N/A	N/A
Conv1d	128	150	1	1	0
MLP layers:					
dropout	N/A	N/A	N/A	N/A	N/A
fc1	300	170	N/A	N/A	N/A
dropout	N/A	N/A	N/A	N/A	N/A
fc2	170	170	N/A	N/A	N/A
dropout	N/A	N/A	N/A	N/A	N/A
fc3	170	82	N/A	N/A	N/A

Table 8: **Object re-recognition model**

	Split 1			Split 2			Split 3		
Max epoch	500			500			500		
Learning rate	3.6515e-05			8.8224e-05			7.8910e-05		
Dropout	0.2348			0.2240			0.3253		
Batch size	200			200			400		
Results with different input:	A+C	A	C	A+C	A	C	A+C	A	C
Best validation accuracy	0.9707	0.8744	0.5295	0.9383	0.8756	0.5105	0.9200	0.8137	0.5456
Best test accuracy	0.9241	0.8034	0.4599	0.9183	0.8281	0.5276	0.9332	0.8920	0.4793

Table 9: **Object re-recognition hyperparameters and results on three random data splits.** In the table, "A" refers to the acoustic vibration signal input modality, and "C" refers to the contact point location input modality.

176 experiments (i.e., shaking and pouring), the duration of the acoustic clip is long, so we chose a larger
177 hop length. We have observed that there are no obvious acoustic signals above the frequency 8192
178 for this experiment, we chose the smaller highest frequency to show more details of the spectro-
179 grams. For rigid object experiments (i.e., dice shape and dice inventory), the duration of collision
180 vibration is very short. To perceive more details of the collision signal, we chose a smaller hop
181 length. Since the collision vibration signal of solid objects includes a higher frequency, we set the
182 highest frequency to be larger.

183 I. Quantitative Results in Characterization Experiments

184 We derived twelve interpretable signal descriptors to quantitatively measure the features of the
185 acoustic vibration signals. These feature descriptors are key statistical summaries of certain aspects
186 of the signals, which provide an interpretable understanding of the signal-capturing capabilities and
187 sensitivity characterization of our robot hand. Specifically, we denote them as (1) D1: average root
188 mean square of the signal, (2) D2: average spectral centroid, (3) D3: average bandwidth, (4) D4:
189 average contrast, (5) D5: average flatness, (6) D6: average roll-off, (7) D7: average zero crossing

parameters	pouring	shaking	dice shape	dice inventory
Length of FFT window	2048	2048	2048	2048
Hop length	2048	1024	512	512
Number of Mel band	64	64	64	64
Highest frequency(Hz)	8192	8192	16384	16384
Color limits	[-10,-80]	[-10,-80]	[-10,-80]	[-10,-80]
Duration s	13.00s	2.38s	3.89s	3.89s
Poly features order	3	3	3	3
Roll-off percentage	0.9	0.9	0.9	0.9

Table 10: **Parameters of acoustic vibration signal processing in characterization experiments.**

190 rate, (8) D8: average tempogram, (9) D9: average poly features, (10) D10: average MFCCs, (11)
191 D11: average chroma, and (12) D12: average tonnetz.

192 We repeated 30 trials for each subtask in the container experiment. In each trial, we extracted the
193 twelve descriptor features by averaging the feature of the acoustic signals from four fingers. We
194 then rescaled all the descriptor values to the range between 0 and 1. With these feature descriptors,
195 we performed unsupervised dimensionality reduction of the high-dimensional signals into 2D space
196 to test whether we can distinguish between various events triggered by different object states. The
197 quantitative results of the mean value and standard error of the mean (SEM) of the twelve acoustic
198 vibration signal descriptors from the 30 trials are shown in Tab. 11.

	D1	D2	D3	D4	D5	D6
Dice(quantity:1)	0.093±0.010	0.249±0.013	0.514±0.032	0.464±0.036	0.182±0.013	0.430±0.023
Dice(quantity:3)	0.590±0.011	0.825±0.016	0.827±0.019	0.3386±0.024	0.782±0.022	0.848±0.015
Dice(quantity:5)	0.803±0.013	0.779±0.023	0.728±0.034	0.306±0.039	0.684±0.034	0.798±0.026
Dice(6 edges)	0.106±0.009	0.839±0.020	0.886±0.009	0.482±0.027	0.613±0.022	0.823±0.016
Dice(12 edges)	0.559±0.011	0.682±0.015	0.655±0.011	0.517±0.035	0.717±0.027	0.615±0.010
Dice(30 edges)	0.802±0.014	0.415±0.027	0.299±0.023	0.695±0.029	0.469±0.036	0.388±0.024
Pouring(1st 100ml)	0.507±0.022	0.386±0.024	0.311±0.025	0.684±0.033	0.360±0.037	0.335±0.024
Pouring(2nd 100ml)	0.821±0.019	0.135±0.0137	0.154±0.015	0.796±0.018	0.326±0.033	0.114±0.010
Pouring(3rd 100ml)	0.751±0.016	0.163±0.0164	0.220±0.16	0.508±0.021	0.400±0.038	0.130±0.011
Shaking(100ml)	0.032±0.004	0.825±0.020	0.853±0.011	0.573±0.052	0.597±0.033	0.919±0.011
Shaking(200ml)	0.219±0.017	0.736±0.018	0.657±0.012	0.582±0.023	0.566±0.023	0.713±0.011
Shaking(300ml)	0.480±0.039	0.318±0.031	0.297±0.026	0.565±0.029	0.231±0.026	0.302±0.027
	D7	D8	D9	D10	D11	D12
Dice(quantity:1)	0.123±0.019	0.161±0.019	0.135±0.012	0.284±0.031	0.298±0.025	0.464±0.048
Dice(quantity:3)	0.739±0.022	0.524±0.021	0.576±0.018	0.281±0.023	0.716±0.020	0.553±0.042
Dice(quantity:5)	0.767±0.025	0.716±0.026	0.782±0.022	0.566±0.027	0.742±0.028	0.587±0.035
Dice(6 edges)	0.524±0.036	0.576±0.031	0.062±0.006	0.188±0.018	0.435±0.033	0.743±0.027
Dice(12 edges)	0.593±0.031	0.203±0.014	0.470±0.008	0.547±0.020	0.486±0.028	0.515±0.025
Dice(30 edges)	0.490±0.038	0.319±0.019	0.820±0.014	0.860±0.016	0.694±0.024	0.359±0.028
Pouring(1st 100ml)	0.441±0.027	0.365±0.045	0.488±0.019	0.624±0.028	0.585±0.040	0.622±0.034
Pouring(2nd 100ml)	0.162±0.016	0.543±0.044	0.818±0.019	0.810±0.018	0.491±0.038	0.567±0.024
Pouring(3rd 100ml)	0.232±0.020	0.469±0.040	0.717±0.017	0.438.016	0.441±0.041	0.235±0.025
Shaking(100ml)	0.630±0.031	0.911±0.013	0.0576±0.006	0.534±0.038	0.532±0.039	0.592±0.047
Shaking(200ml)	0.706±0.031	0.260±0.013	0.236±0.011	0.642±0.011	0.678±0.032	0.345±0.032
Shaking(300ml)	0.295±0.034	0.112±0.009	0.606±0.032	0.842±0.012	0.478±0.027	0.568±0.04

Table 11: **Quantitative results of the mean value and standard error of the mean (SEM) of the twelve acoustic vibration signal descriptors in characterization experiments.**

199 **References**

- 200 [1] E. Crossley, T. Biggs, P. Brown, and T. Singh. The accuracy of iphone applications to monitor
201 environmental noise levels. *The Laryngoscope*, 131(1):E59–E62, 2021.