2010 IEEE International Conference on Robotics and Automation
Anchorage Convention District
May 3-8, 2010, Anchorage, Alaska, USA

How to Separate Containers From Non-Containers? A Behavior-Grounded Approach to Acoustic Object Categorization

Shane Griffith, Jivko Sinapov, Vladimir Sukhoy, and Alexander Stoytchev
Developmental Robotics Laboratory
Iowa State University
{shaneg, jsinapov, sukhoy, alexs}@iastate.edu

***Abstract*— This paper describes an approach to interactive object categorization that couples exploratory behaviors and their resulting acoustic signatures to form object categories. The framework was tested with an upper-torso humanoid robot on a container/non-container categorization task. The robot used six exploratory behaviors (drop block, grasp, move, shake, flip, and drop object) and applied them to twenty objects. The results from this large-scale experimental study show that the robot was able to learn meaningful object categories using only acoustic information. The results also show that the quality of the categorization depends on the exploratory behavior used to derive it as some behaviors elicit more salient acoustic signatures than others.**

I. INTRODUCTION

The ability to form meaningful object categories is one of the hallmarks of human infant development [1]. Infants as young as 6-months-old can learn an abstract representation of a simple object category [2]. Furthermore, theories in psychology and cognitive science have proposed that active interaction with objects is necessary to form categories that capture the functional properties of an object [3]. Traditionally, however, most methods for object recognition and object categorization have been vision-based (see [4] for a literature review). Because these methods rely on passive observation (as opposed to active exploration) they often fail to capture the functional properties of objects. For example, two objects that look the same are indistinguishable with vision but they may produce different sounds when shaken. Similarly, it is very hard to specify what a container looks like, but it may be very easy to detect a container by dropping an object over it and listening for the specific sound pattern of the object bouncing inside the container.

In contrast to disembodied vision-based systems, humans and many animals use active behavioral exploration to learn about and to classify novel objects [5]. Furthermore, humans ground object knowledge using multiple modalities (e.g., touch and hearing) in addition to vision. Similar behavior-grounded approaches have proven quite useful in robotics as well [6] [7]. The advantage of using behaviors to ground object information is that the robot can autonomously test, verify, and correct its own knowledge representation without human intervention [8] [9].

A growing body of empirical studies in embodied acoustic object recognition supports this view [10] [11] [12] [13] [14]. These studies have shown that probing an object and other forms of simple contact are sufficient for a robot to identify the material type from which the object is made of. A robot can become better at object recognition as it performs more exploratory behaviors on an object [13]. Further work is necessary, however, to determine if a robot can use similar acoustic models to form object categories.

Fig. 1. The upper-torso humanoid robot used in the experiments. The robot is shaking one of the container objects used in the experiments.

This paper tests the assumption that a robot can form meaningful categories of objects using only acoustic information. The robot's task was to categorize 10 container and 10 non-container objects using six different exploratory behaviors (*drop block*, *grasp*, *move*, *shake*, *flip*, and *drop object*). The robot automatically formed acoustic outcome classes by clustering the sounds it observed for a given behavior. Object categories were determined using the frequency with which different acoustic outcomes occurred with different objects. The results show that the robot was indeed able to form meaningful object categories. The results also show that the number of interactions and the choice of exploratory behavior affect the quality of the categorization as some behaviors are better suited for this task than others.

II. RELATED WORK

Relatively few studies have investigated how a robot can *ground* the representation of object categories in its sensorimotor experience. Perhaps the first work toward interactive object categorization was done by Pfeifer and Scheier [15], in which a mobile robot traversed its environment with the task of cleaning it. The robot could lift small objects and push

medium-sized objects. The robot learned to collect small and medium-sized objects and to ignore large objects.

Several studies have shown how a robot can learn similarities among different types of objects. The robot in the work described by Sinapov and Stoytchev [16] interacted with 6 different stick-shaped tools and learned a hierarchical taxonomy of outcomes for each one. It computed the functional similarity between two tools by comparing their outcome taxonomies. In another study, Montesano *et al.* [17] created a framework with which a robot learned the similarity between differently sized spheres and cubes by learning relationships between the robot's interactions, the object's features, and the observed effects. In the work of Ugur *et al.* [18], a simulated robot traversed an environment with random dispersions of spheres, cubes, and cylinders. It learned which objects afforded traversability (spheres and cylinders in lying orientations) from those that did not (cubes and cylinders in upright orientations). None of the robots in [16], [17], or [18] performed explicit object categorization.

Metta and Fitzpatrick [19] [6] showed that a robot could simplify the task of object segmentation and recognition by probing its environment. When the robot's arm made contact with an object it detected a unified area of movement that it used to delineate the object from the background. This helped the robot learn an object model for recognition. The robot also observed the different movement outcomes for each object (e.g., *rollable* and *non-rollable*), which it associated with the object model.

In the work of Nakamura *et al.* [20] a robot captured multimodal object data, which was used to infer the object properties in one modality using data from another. The robot squeezed objects to capture hardness, shook objects to capture sound, and viewed objects from different angles to capture visual features. They showed that the robot could infer the hardness of an object from visual information much better than it could infer whether the object would make noise using visual information.

In the work of Sinapov and Stoytchev [13], a robot recognized objects using only acoustic data. The robot acquired an interaction history of 1800 behaviors by performing 5 interactions (*grasp*, *shake*, *drop*, *push*, and *tap*) 10 times on 36 objects. The robot was able to recognize objects from novel acoustic outcomes with 73% accuracy. The recognition accuracy increased to 99% when the robot was allowed to perform all 5 behaviors on the object before determining its identity. In a follow-up study [14], the robot was also able to classify objects based on their material type.

Sahai *et al.* [21] used a robot to categorize 12 different objects and 12 different surfaces. The categories captured differences in the usefulness of objects and surfaces for robot writing tasks. The robot detected marks as it performed 10 trace-making behaviors with each object-surface pair. The robot categorized objects using the frequency with which each object left a mark on each surface. It categorized surfaces using the frequency with which each surface captured the traces left by each object.

In our previous work [22], a robot categorized 5 containers and 5 non-containers using visual information. The robot dropped a block over an object and observed co-movement patterns between the block and the object as it pushed the object. It formed outcome classes by clustering its observations of co-movement. It formed object categories by clustering the objects based on the frequency with which different co-movement outcomes occurred with each object. The separation of containers and non-containers allowed the robot to learn a visual representation of each category from 3D depth images, which it used to quickly identify the category of novel objects.

This paper builds on our previous work [22] by adding more exploratory behaviors (now 6 instead of 1), increasing the number of the behavioral interactions with the objects (now 12000 instead of 1000), and capturing acoustic data instead of visual movement data. In [22] the robot learned the object categories using visual co-movement features specified by a human. In this paper the robot automatically extracted acoustic features, after exploring 20 objects, and learned from these features in an unsupervised way. It should be noted that in this paper, the identity of each object is assumed to be known. In other words, the acoustic data corresponding to actions on a specific object is labeled with the object ID. What is unlabeled is the category (container versus non-container).

III. EXPERIMENTAL SETUP

A. Robot

All experiments were performed with the upper-torso humanoid robot shown in Fig. 1. The robot was built with two 7-DOF Whole Arm Manipulators (WAMs) by Barrett Technology, each equipped with the Barrett Hand as its end effector. The WAMs are mounted in a configuration similar to that of human arms. They are controlled in real time from a Linux PC at 500 Hz over a CAN bus interface.

The audio data for the experiments was collected with an Audio-Technica U853AW UniPoint Cardioid Condenser Hanging Microphone mounted in the robot's head. The microphone's output was routed through an ART Tube MP Studio Microphone pre-amplifier and a Lexicon Alpha bus-powered interface, which transmits sound to the PC via USB. Audio was recorded at 44.1 KHz over a 16-bit channel using the Java Sound API.

B. Objects

The robot interacted with a small plastic block and 10 different objects (shown in Fig. 2). Each of the 10 objects was a container in one orientation and a non-container when flipped over. Flipping the containers was an easy way for the robot to learn about non-containers while preserving the dimensions of the objects in the two categories.

The objects were selected to have a variety of shapes, sizes, and materials. Objects were tall, short, rectangular and round. They were made of plastic, metal, wicker, and foam. A few objects that were initially selected could not be

Fig. 2. The objects used in the experiments. (**Containers**) The first two rows show the 10 container objects: wicker basket, metal trash can, potpourri basket, flower pot, bed riser, purple bucket, styrofoam bucket, car trash can, green bucket, and red bucket. (**Non-containers**) The second two rows show the same 10 objects as before but flipped upside down, which makes them non-containers for this particular robot with this particular set of behaviors.

used because they were too large to be grasped. Also, the aluminum fingers of the Barrett Hand did not create a firm grip with many objects, which was important for a large-scale experimental study like this one. Therefore, rubber fingers were stretched over each of the robot's three fingers to achieve more reliable grasps.

C. Robot Behaviors

Six behaviors were performed during each trial: 1) *drop* the block, 2) *grasp* the object, 3) *move* the object, 4) *shake* the object, 5) *flip* the object, and 6) *drop* the object. A person placed the block and the object at specific locations before the start of each experiment. The robot grasped the block and positioned its hand in the area above the object before executing the six behaviors listed above. Figure 3 shows the sequence of interactions for two separate trials.

The drop positions for the *drop block* behavior were randomly selected from a 2D Gaussian distribution centered above the object. The standard deviation was empirically set to be equal to the width (in pixels) of each object. Thus, the small block fell inside a container during approximately 70% of all trials with containers. During the other 30% of the trials with containers (and during trials with non-containers) the block fell on the table. In some cases the block rolled off the table (approximately 5% of all trials). In these cases, the block was left off the table for the duration of the trial.

The other behaviors are self-explanatory (see Fig. 3).

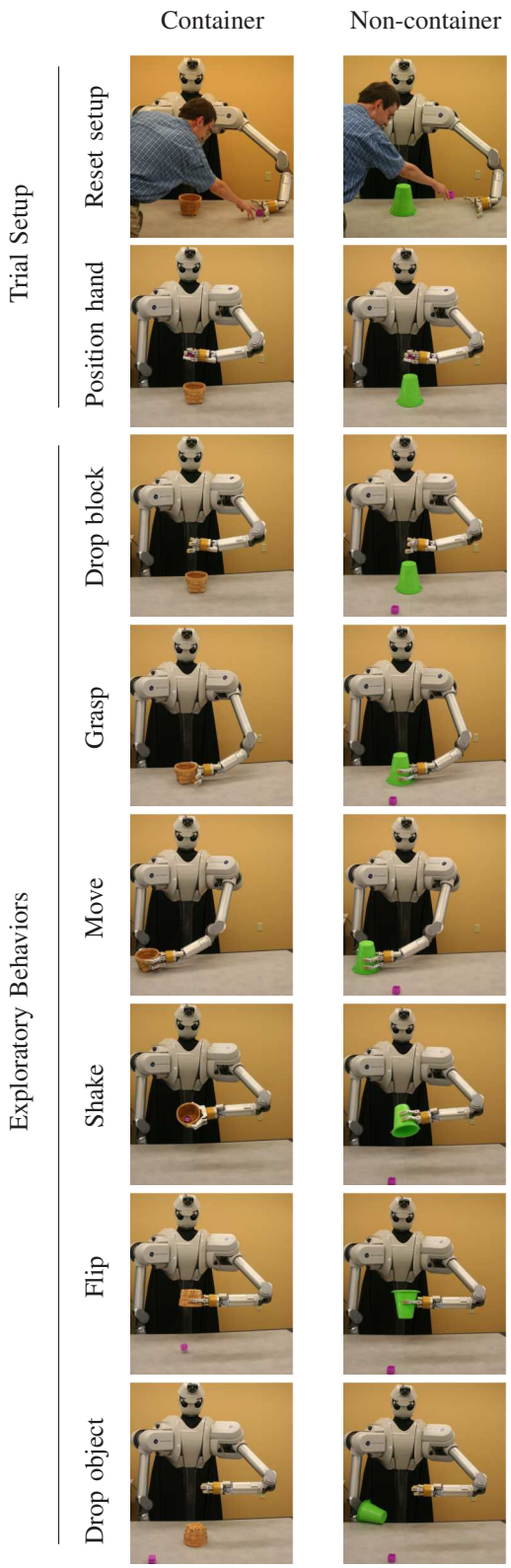

Fig. 3. Snapshots from two separate trials with a container and a non-container object. Before each trial a human experimenter reset the setup by placing the block and the object at marked locations. After grasping the block and positioning its arm at a random location above the object the robot performed the six exploratory behaviors one after another.

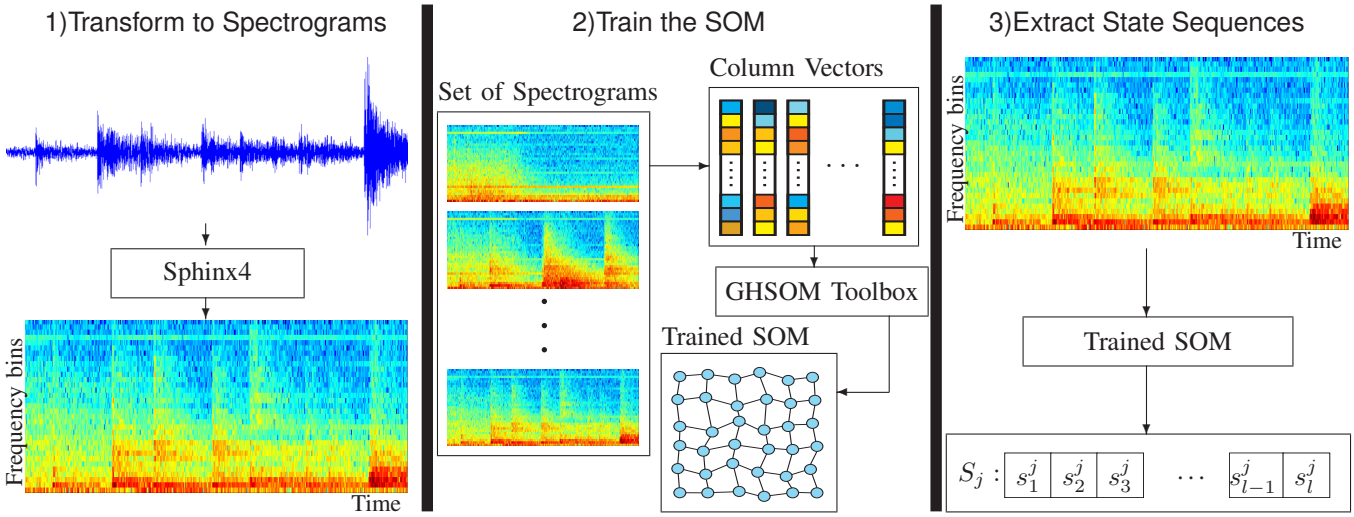

Fig. 4. The feature extraction process: 1) The raw sound wave produced by each behavior is transformed to a spectrogram. Each spectrogram has 33 bins (represented as column vectors), which capture the intensity of the audio signal for different frequencies at a given time slice. Red color indicates high intensity while blue color indicates low intensity. 2) An SOM is trained using randomly selected column vectors from the spectrograms for a given behavior. 3) The column vectors of each spectrogram are mapped to a discrete state sequence using the states of the SOM. Each column vector is mapped to the most highly activated SOM node when the column vector is used as an input to the SOM. See the text for more details.

IV. METHODOLOGY

A. Data Collection

The robot collected multiple audio sequences while performing each of the six exploratory behaviors, $\mathcal{B}$ = [*drop block*, *grasp*, *move*, *shake*, *flip*, *drop object*]. The six behaviors were organized into trials and always performed one after another (see Fig. 3). For each of the 20 objects (10 containers and 10 non-containers) the robot performed 100 trials, for a total of $20 \times 100 = 2000$ trials. With 6 behaviors per trial, the robot performed $6 \times 2000 = 12000$ behavioral interactions.

Another way to describe this dataset is to say that each behavior (e.g., *shake*) was performed 100 times on each of the 20 objects. Thus, each of the six behaviors was performed 2000 times. During every interaction the tuple (B, O, A) was recorded, where $B \in \mathcal{B}$ was one of the six behaviors performed on object $O \in \mathcal{O}$, and A was the recorded audio sequence.

To minimize the effect of changing background noise while collecting a dataset of this magnitude, the robot performed one trial with each of the twenty objects shown in Fig. 2 before moving on to the second trial with the first object, and so on. This order was chosen to keep slow changes in background noise (e.g., air-conditioning and computer fans) decorrelated from other variables such as object identity or test behavior.

B. Feature Extraction

Auditory features were extracted automatically by representing the sounds produced by each behavioral interaction as a sequence of nodes in a Self-Organizing Map (SOM). The feature extraction process is the same as in our previous work [13]. The three stage process includes: 1) a Discrete Fourier Transform which takes a 44.1 KHz audio sample, A^i, and converts it to a 33 bin spectrogram, $P_i = [p_1^i, \ldots, p_l^i]$, where $p_j^i \in \mathbb{R}^{33}$; 2) a 2D SOM that is trained with the spectrograms corresponding to one of the robot's six exploratory behaviors; and 3) a mapping, $\mathcal{M}(p_j^i) \rightarrow s_j^i$, of each spectrogram column vector, p_j^i, to the most highly activated state, s_j^i, in the SOM when p_j^i is presented as an input to the SOM (see Fig. 4). The mapping process results in a state sequence $S_i = s_1^i s_2^i \ldots s_l^i$, where each s_j^i stands for one of the SOM nodes.

The robot performed this procedure six times, once for every behavior. It acquired a set of state sequences, $\{S_j\}_{j=1}^{2000}$, for each of its six behaviors. This feature extraction method was chosen because it does not require a human to select the acoustic features that the robot will have to use. The algorithm identifies and computes features in an unsupervised way. See [13] for further details.

C. Learning Auditory Outcome Classes

The acoustic outcome patterns produced by a given behavior can be clustered automatically to obtain auditory outcome classes. As the number of interactions increases, the learned outcome classes gradually become more stable and more robust to outliers (see section V.C). In our case, the robot's task was to learn 6 separate sets of acoustic outcome classes—one for each behavior. More formally, the robot learned k outcome classes $C = \{c_1, ..., c_k\}$ from the set of SOM state sequences, $\{S_j\}_{j=1}^{2000}$, observed during the execution of one of the 6 behaviors. An unsupervised hierarchical clustering procedure based on the *spectral clustering* algorithm was used for this task (spectral clustering is a similarity-based clustering algorithm [23]). The procedure was performed 6 different times to obtain 6 different sets of acoustic outcome classes. Figure 5 illustrates the process for only one of them.

The *spectral clustering* algorithm requires a similarity matrix as its input. The similarity between acoustic outcomes,

S_a and S_b, represented as sequences of SOM states produced by two different executions of the same behavior was determined using the Needleman-Wunsch global alignment algorithm [24] [25]. The algorithm can estimate the similarity between any two sequences if the data is represented as a sequence over a finite alphabet. The general applicability of the algorithm has made it popular for other applications such as comparing biological sequences, text sequences, and more [25]. Computing the similarity of two sequences requires a substitution cost (i.e., a difference function) to be defined for any two tokens in the finite alphabet. Here the substitution cost is defined as the Euclidean distance between any two nodes in the SOM (each node in the 2D SOM has an x and a y coordinate).

The resulting similarity matrix, $\mathbf{W}$, was used as input to the unsupervised hierarchical clustering procedure, which partitions the input data points (i.e., audio sequences) into disjoint clusters. The spectral clustering algorithm exploits the eigenstructure of the matrix to partition the data points. Finding the optimal graph partition is an NP-complete problem. Therefore, the Shi and Malik [26] approximation algorithm was used, which minimizes the *normalized cut* objective function. The following steps give a summary of the algorithm:

1) Let $\mathbf{W}_{n\times n}$ be the symmetric matrix containing the similarity score for each pair of acoustic outcome sequences.
2) Let $\mathbf{D}_{n\times n}$ be the degree matrix of $\mathbf{W}$, i.e., a diagonal matrix such that $\mathbf{D}_{ii} = \sum_j W_{ij}$.
3) Solve the eigenvalue system $(\mathbf{D} - \mathbf{W})x = \lambda \mathbf{D}x$ for the eigenvector corresponding to the second smallest eigenvalue.
4) Search for a threshold of the resulting eigenvector to create a bi-partition of the set of acoustic outcomes that minimizes the normalized cut objective function. Accept this bi-partition if the resulting value of the objective function is smaller than a threshold α.
5) Recursively bi-partition subgraphs obtained in step 4 that have at least β acoustic sequences.

The output of this procedure is k classes of acoustic outcomes $C = \{c_1, ..., c_k\}$, which are represented as the leaf nodes in a tree structure (see Fig. 5). In our previous work [14], the value for α used in step 4 was set to 0.995. The same value was used here as well. The value for β used in step 5 was empirically set to 40% of the size of the dataset that was initially passed to the spectral clustering algorithm.

D. Object Representation and Categorization

The frequency with which some acoustic outcomes occur with different objects can be used to cluster the objects into categories. For example, when the robot drops a block over a container, it will hear the sound of the block bouncing inside the container more often than when it drops the block over a non-container, in which case it falls on the table.

Given a set of **acoustic outcome classes** $C = \{c_1, \ldots, c_k\}$ extracted from multiple behavioral interactions with objects

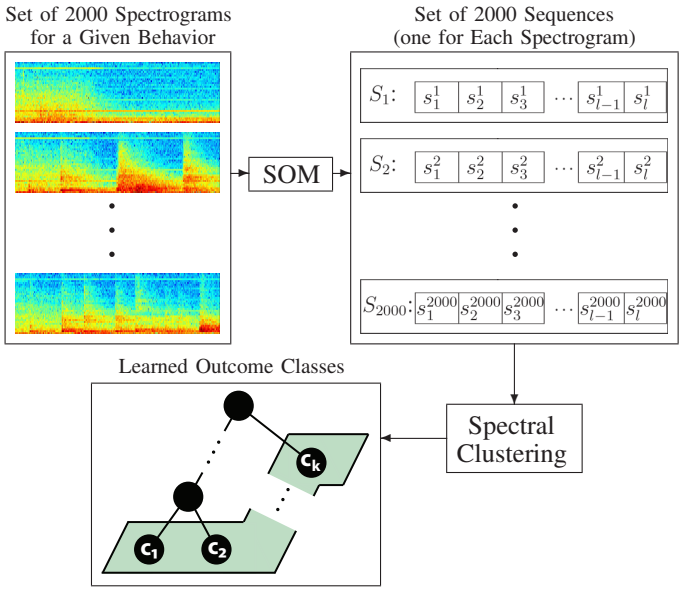

Fig. 5. Illustration of the process used to learn acoustic outcome classes. Each spectrogram is transformed into a state sequence using the trained SOM, which results in 2000 sequences, $\{S_j\}_{j=1}^{2000}$, for each behavior. The acoustic outcome classes are learned by recursively applying the spectral clustering algorithm on this set of sequences. Acoustic outcomes, $C = \{c_1, ..., c_k\}$, are the leaf nodes of the tree created by the recursive algorithm.

$\mathcal{O} = \{O_1, \ldots, O_{20}\}$, the robot computed an outcome occurrence vector $H_v = [h_1^v, \ldots, h_k^v]$ for each object O_v. The value of each h_j^v represents the number of times the acoustic outcome c_j occurred with object O_v, divided by the total number of interactions (100 interactions in this case). In other words, each outcome occurrence vector H_v encodes a probability distribution over the set of outcome classes, such that h_j^v estimates the probability of observing outcome class c_j with object O_v over the entire history of interactions.

The robot formed **object classes** by clustering the feature vectors $H_1, \ldots, H_{20}$ (one for each of the 20 objects shown in Fig. 2). The X-means unsupervised clustering algorithm was used for the procedure. X-means extends the standard K-means algorithm to automatically estimate the correct number of clusters in the dataset [27]. The robot used this strategy to categorize the objects. Six different categorizations were constructed, one for each of the six exploratory behaviors. The results are described in the next section.

V. RESULTS

A. Object Categorization

Four of the six behaviors produced acoustic signals that could be used for object categorization: *drop block*, *shake*, *flip*, and *drop object*. The (mostly silent) *grasp* and *move* behaviors produced acoustic signals that were very similar for all objects and the algorithm clustered all 20 objects into the same object class. Therefore, the results for these two behaviors are not discussed any further. Figure 6 visualizes the categorizations produced by the other four behaviors.

The *drop block* behavior produced three clusters that were almost homogeneous. One cluster had containers and the tall metal non-container (the only misclassified object); one cluster had the rest of the non-containers; and one cluster had the three soft material container baskets. The difference between the softness and hardness of the objects' materials was distinctive enough to create two container categories (cluster 1 and cluster 3 in Fig. 6). The two wicker baskets and the styrofoam bucket are made of soft materials which muffled the block's sound. When the block fell into a hard container it bounced around longer and produced a louder sound.

The *shake* behavior produced results similar to the drop block behavior. In this case, however, there were only two clusters and the three soft material container baskets were incorrectly classified as non-containers. These three objects produced very little sound when shaken, even if the block was inside them. Hence, they sounded similar to the non-containers, which seldom made noise during this interaction. The tall metal non-container was again misclassified.

The *flip* behavior was the most reliable way to discriminate between containers and non-containers in our experiments. It produced a perfect classification. Flipping the object over produced a distinct sound in the case of containers as the small block fell onto the table. In the case of non-containers, no sound was generated as the block was already on the table.

The *drop object* behavior resulted in clusters that were completely heterogeneous. The behavior did not produce different acoustic outcomes for containers and non-containers.

B. Evaluating the Categorization using Information Gain

The category information gain was computed in order to check whether the robot was able to extract meaningful object clusters. The information gain captures how well the object categories formed by the robot resemble the categories specified by a human. The information gain is high when the category labels assigned to the objects match human-provided category labels. It is low otherwise. In other words, if the information gain is high, then the robot has categorized the objects in a meaningful way (even though the robot does not know the human words corresponding to the categories).

Let $\lambda^{(f)} = [\mathcal{O}^1 \dots \mathcal{O}^{M_f}]$ define an object categorization for behavior B_f, where $\mathcal{O}^i$ is the set of objects in the i^{th} cluster. Let p_c^i and p_{nc}^i be the estimated probability that an object drawn from the subset $\mathcal{O}^i$ will be a container or a non-container, respectively. Given a cluster of objects $\mathcal{O}^i$, the Shannon entropy of the cluster is defined as:

$$\mathcal{H}(\mathcal{O}^i) = -p_c^i log_2(p_c^i) - p_{nc}^i log_2(p_{nc}^i)$$

In other words, an object cluster containing mostly containers or mostly non-containers will have low entropy, while a cluster containing an equal number of containers and non-containers will have the maximum entropy. Hence, the information gain for the entire object categorization $\lambda^{(f)} = [\mathcal{O}^1 \dots \mathcal{O}^{M_f}]$, learned using behavior B_f, is given by the following formula:

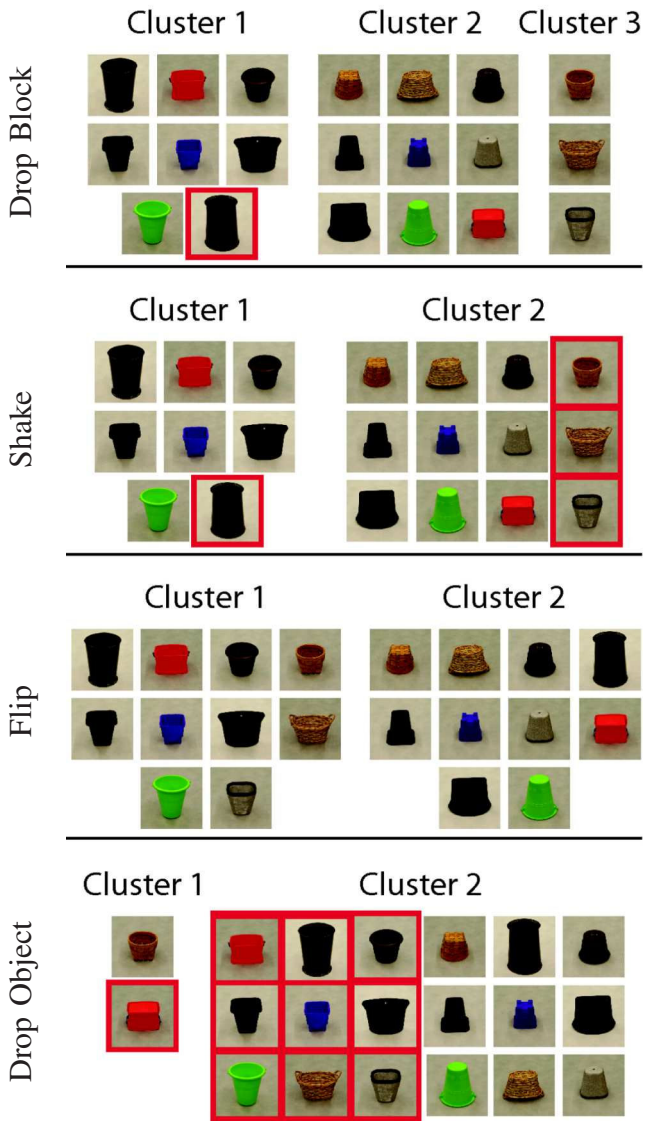

Fig. 6. Visualization of the object categories formed by the robot for four of the six exploratory behaviors. The quality of the classification depends on the behavior that was performed. The *flip* behavior produced a perfect classification. The *grasp* and *move* behaviors both produced only one cluster with all twenty objects in it so their results are not visualized. The other behaviors produced clusters that were not always so pure. Incorrect classifications (determined from ground truth category labels provided by a human and the majority class of the cluster) are framed in red.

$$IG(\lambda^{(f)}) = \mathcal{H}(\mathcal{O}) - \sum_{i=1}^{M_f} \frac{|\mathcal{O}^i|}{|\mathcal{O}|} \mathcal{H}(\mathcal{O}^i)$$

To get a baseline information gain value for comparison, the information gain was computed for a random labeling. That is, the values for p_c^i and p_{nc}^i were estimated after randomly shuffling the labels of the objects in the clusters $\mathcal{O}^i$ (for $i = 1$ to M_f) while preserving the number of objects in each cluster. The procedure was repeated 100 times to estimate the mean and the standard deviation. Figure 7 shows the information gain for each categorization and compares it to the corresponding baseline average random information gain.

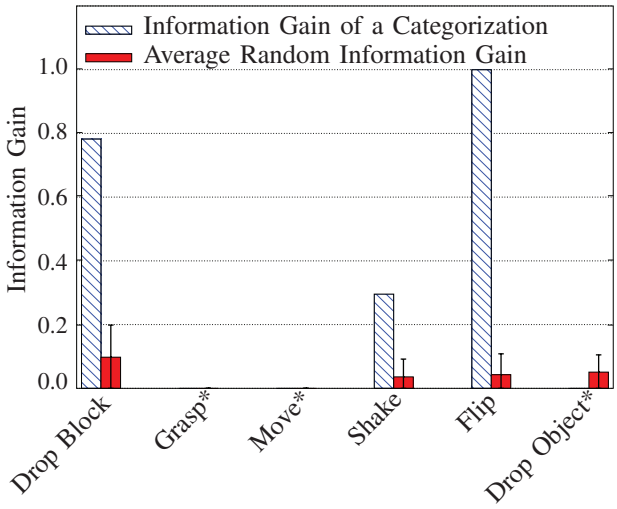

Fig. 7. Information gain for the categorizations formed by each of the 6 behaviors after the robot has performed 100 interactions with each of the 20 objects. For comparison, the figure also shows the average information gain for a random categorization, which was computed by shuffling the category labels of the objects in each cluster 100 times and estimating the mean and the standard deviation of the information gain. Human-provided labels were used for this evaluation procedure (these labels were not used in any part of the robot's learning process). Three behaviors had an information gain of zero, which is denoted by the * character.

The figure shows that the *flip* and the *drop block* behaviors have the highest information gain with respect to the average random labeling. The information gain for *shake* shows that it performed significantly better than chance, albeit not as well as we expected. The remaining three behaviors had zero information gain, illustrating that they did not produce meaningful categorizations. These results show that some behaviors can be used to form meaningful object categories. The next section shows how the number of interactions with each object affects the quality of a categorization.

C. Categorization Performance vs. Number of Interactions

The number of behavioral interactions used by the categorization procedure was varied to determine if the quality of a categorization improves when more interactions are performed. Presumably, for behaviors that have information gain greater than zero the quality of the categorization with respect to human labels would improve as the number of interactions increases. This section tests this hypothesis.

The evaluation was performed by randomly sampling smaller datasets from the larger dataset described above. More specifically, N interactions were sampled at random from the 100 interactions performed with each of the 20 objects. A new categorization was formed from this new dataset by: 1) re-training the SOM; 2) converting the spectrograms to state sequences; 3) forming outcome classes from the set of state sequences using spectral clustering; and 4) categorizing objects by their acoustic outcome frequencies. The quality of the categorization was determined using the information gain formula described in the previous section. The process was repeated 10 times for each value of N in order to estimate the mean and standard deviation of the information gain. Figure 8 shows the results.

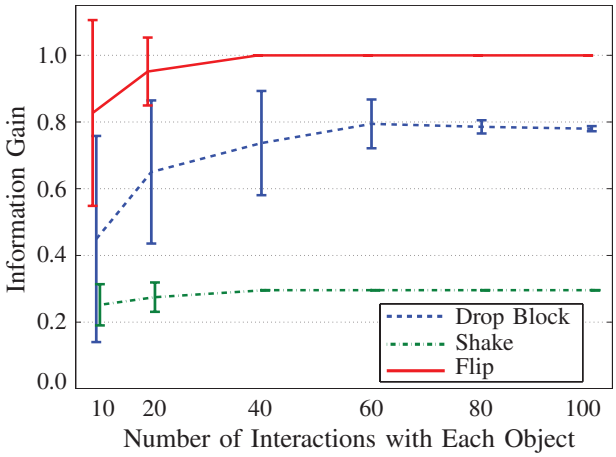

Fig. 8. Information gain for the categorizations formed by the *drop block*, *shake*, and *flip* behaviors as the number of interactions with each object is increased. This graph was computed by randomly sampling N interactions from the 100 interactions with each object and re-running the learning algorithms on the smaller dataset. This process was repeated 10 times for each value of N to estimate the mean and standard deviation. Human-provided category labels were used to compute the information gain.

This experiment was performed only for the three behaviors that produced categorizations with non-zero information gain (*drop block*, *shake*, or *flip*) using the entire dataset (see Fig. 7). Figure 8 shows that the information gain of the resulting categorizations first increases and then converges after only 40 interactions. The *drop block* behavior requires 60 interactions. The figure also shows that the variance of the information gain with respect to human labeling converges to zero as the number of interactions increases. This is true for all three behaviors.

VI. CONCLUSIONS AND FUTURE WORK

This paper described a framework that allowed a robot to interactively categorize objects based on the acoustic outcomes that they produce when the robot applies different exploratory behaviors on the objects. The framework is based on the idea that knowledge about objects should be grounded in the behavioral and perceptual repertoire of the robot [8] [9]. A large-scale experimental study with an upper-torso humanoid robot was conducted to evaluate this framework. A container/non-container categorization task with 20 objects was chosen for this evaluation. The fact that meaningful categories were produced with so many objects lends further credence to the hypothesis that a robot can interactively categorize objects using the frequency with which different perceptual outcomes occur with each object.

The results demonstrate that the categorization accuracy is highly dependent on the behavior that the robot used to produce the categorization. Some behaviors simply capture the 'container' property better than others. Interestingly, the behaviors that best discriminated between containers and non-containers caused the block to become contained (which occurred during the *drop block* behavior) and to become uncontained (which occurred during the *flip* behavior). The *drop object* behavior did not produce outcomes specific to the container object category. This suggests that the inter-

active behaviors that can best discriminate between object categories are behaviors that capture some category-specific property. Indeed, the results show that the robot performed well when category-specific interactions were used.

It was also shown that the robot can split the objects into meaningful categories even though it does not know the mapping between these categories and the human words for them. What the robot does know, however, is that the objects in a given category produce similar distributions of acoustic outcomes. The robot also knows that the differences between categories can be explained in terms of the frequencies of the detected acoustic events.

The are several possible directions for future work. For example, the framework described here performed well with data from a single behavior and a single sensory modality (audio). It would be desirable to investigate how a robot can combine its observations from executions of different behaviors to come up with a single, unified object categorization, instead of one separate categorization for each behavior.

Future work should also examine how a robot can learn object categories without using explicit object IDs. Another direction for future work is to investigate how to combine observations from multiple modalities (e.g., vision and audio). Combining information from multiple modalities is useful because one modality may capture discriminative information that another modality may miss. For example, while vision can discriminate between containers and non-containers using the *move* behavior [22], audio cannot.

REFERENCES

[1] L. Cohen, "Unresolved issues in infant categorization," in *Early category and concept development*, D. Rakison and L. M. Oakes, Eds. New York: Oxford University Press, 2003, pp. 193–209.

[2] M. Casasola, L. B. Cohen, and E. Chiarello, "Six-month-old infants' categorization of containment spatial relations," *Child Development*, vol. 74, no. 3, pp. 679–693, May 2003.

[3] J. Mandler, "Thought before language," *Trends in Cognitive Sciences*, vol. 8, no. 11, pp. 508–513, November 2004.

[4] A. Pinz, "Object categorization," *Foundations and Trends in Computer Graphics and Vision*, vol. 1, no. 4, pp. 255–353, Dec 2005.

[5] T. Power, *Play and Exploration in Children and Animals*. Mahwah, NJ: Laurence Erlbaum Associates, 2000.

[6] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini, "Learning about objects through action - initial steps towards artificial cognition," in *Proc. of the 2003 IEEE Intl. Conf. on Robotics and Automation*, 2003, pp. 3140–3145.

[7] A. Stoytchev, "Behavior-grounded representation of tool affordances," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2005, pp. 3071–3076.

[8] R. Sutton, "Verification, the key to AI," on-line essay. [Online]. Available: http://www.cs.ualberta.ca/ sutton/IncIdeas/KeytoAI.html

[9] A. Stoytchev, "Some basic principles of developmental robotics," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 2, pp. 122–130, 2009.

[10] E. Krotkov, R. L. Klatzky, and N. B. Zumel, "Robotic perception of material: Experiments with shape-invariant acoustic measures of material type," in *Proc. of the 4th International Symposium on Experimental Robotics IV*. London, UK: Springer-Verlag, 1997, pp. 204–211.

[11] K. Richmond and D. Pai, "Active measurement of contact sound," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation*, 2000, pp. 2146–2152.

[12] E. Torres-Jara, L. Natale, and P. Fitspatrick, "Tapping into touch," in *Proc. of the Fifth Intl Workshop on Epigenetic Robotics*, 2005, pp. 79–86.

[13] J. Sinapov, M. Wiemer, and A. Stoytchev, "Interactive learning of the acoustic properties of household objects," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, Kobe, Japan, May 2009, pp. 3937–3943.

[14] J. Sinapov and A. Stoytchev, "From acoustic object recognition to object categorization by a humanoid robot," in *Proc. of the RSS 2009 Workshop - Mobile Manipulation in Human Environments*, Seattle, WA, June 2009.

[15] R. Pfeifer and C. Scheier, "Sensory-motor coordination: The metaphor and beyond," in *Robotics and Autonomous Systems*, vol. 20, 1997, pp. 157–178.

[16] J. Sinapov and A. Stoytchev, "Detecting the functional similarities between tools using a hierarchical representation of outcomes," in *Proceedings of the 7th IEEE International Conference on Development and Learning*, Monterey, CA, August 2008.

[17] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: From sensory-motor coordination to imitation," *IEEE Trans. on Robotics*, vol. 24, no. 1, pp. 15–26, 2008.

[18] E. Ugur, M. Dogar, M. Cakmak, and E. Sahin, "The learning and use of traversability affordance using range images on a mobile robot," in *Proc. of the IEEE Intl Conf. on Robotics and Automation*, 2007, pp. 1721–1726.

[19] G. Metta and P. Fitzpatrick, "Early integration of vision and manipulation," *Adaptive Behavior*, vol. 11, no. 2, pp. 109–128, June 2003.

[20] T. Nakamura, T. Nagai, and N. Iwahashi, "Multimodal object categorization by a robot," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2007, pp. 2415–2420.

[21] R. Sahai, S. Griffith, and A. Stoytchev, "Interactive identification of writing instruments and writable surfaces by a robot," in *Proc. of the RSS 2009 Workshop - Mobile Manipulation in Human Environments*, Seattle, WA, June 2009.

[22] S. Griffith, J. Sinapov, M. Miller, and A. Stoytchev, "Toward interactive learning of object categories by a robot: A case study with container and non-container objects," in *Proc. of the 8th IEEE Intl. Conf. on Development and Learning (ICDL)*, Shanghai, China, June 2009.

[23] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[24] S. Needleman and C. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, vol. 48, no. 3, pp. 443–453, 1970.

[25] G. Navarro, "A guided tour to approximate string matching," *ACM Computing Surveys*, vol. 33, no. 1, pp. 31–88, 2001.

[26] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[27] D. Pelleg and A. W. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," in *Proc. of the Seventeenth International Conference on Machine Learning*, 2000, pp. 727–734.