Speech Enhanced Imitation Learning and Task Abstraction for Human-Robot Interaction

Simon Stepputtis[1], Chitta Baral[1], Heni Ben Amor[1]

***Abstract*— In this short paper, we show how to learn interaction primitives and networks from long interactions by taking advantage of language and speech markers. The speech markers are obtained from free speech that accompanies the demonstration. We perform experiments to show the value of using speech markers for learning interaction primitives.**

I. Introduction

The broader goal of our research is to facilitate imitation learning using language and speech markers. In this short paper, we show how to learn Interaction Primitives (IP) [1] and networks from long interactions by taking advantage of speech markers.

As collaborative robots become increasingly available, methodologies and tools are needed that allow them to expand their repertoire of interaction skills. Programming such skills by hand is a challenging endeavor since it requires anticipating and a-priori reasoning about the situations that may occur. While imitation learning [2], [3], [4] can be used to facilitate this process, there are many important aspects of a collaborative task that cannot be communicated through behavioral demonstrations only, e.g., the individual segments of the task, the semantic type of behavior executed, or the name of the target object. Indeed, human teachers and coaches often use a combination of motion and language to convey a variety of information to a student. Consequently, novel imitation learning approaches are needed that leverage both modalities.

In this paper, we investigate how verbal instructions extracted from human speech can be used to segment and semantically annotate human demonstrations. Furthermore, we show that this information can be used to learn both (a) low-level interaction primitives, as well as (b) higher-level interaction networks that encode the transition model among primitives. As a result, few(er) demonstrations are necessary to learn both the motion and structure underlying the imitated task. In addition, the recorded human speech markers also provide information about semantic aspects of the task, e.g., synonyms are mapped onto the same internal representation.

Although there has been work on using human language to teach robots [5], [6], [7], [8], [9], [10], the majority of these approaches focuses on language-only instruction modes. Our work is similar in spirit to the work in Akgun et al. [11]. However, we use free speech to outline one of multiple objects as object of interest and learn from long interactions, thus being able to learn multiple skills from a single demonstration.

[1]Simon Stepputtis, Chitta Baral and Heni Ben Amor are with the School of Computing, Informatics and Decision Systems Engineering, Arizona State University, 660 S. Mill Ave, Tempe, AZ 85281 {ssteppu t , chitta , hbenamor} at asu.edu

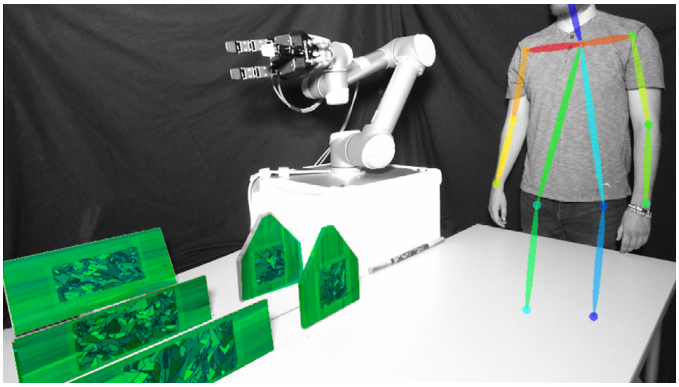

Fig. 1. This figure shows the environment of the robot as well as its initial setup for the experiments. It combines the results from the object tracker (Simtrack) which highlights the tracked objects in green and the skeleton tracker (OpenPose) which draws a skeleton on the limbs of every detected human.

II. Methodology

We explain our methodology to learn interaction primitives from a few long interactions with respect to the following scenario shown in Figures 1 and 2. Figure 1 shows the environment as perceived by the robot. Multiple objects as well as the humans skeleton are tracked and thus, can be recognized by the robot. The image sequence in Figure 2 shows the assembly of a wooden toolbox where each part (except the handle) is out of reach from the human. The task of the robot is to assist the human by handing over all parts, such that the toolbox can be assembled successfully.

Our research goal is to observe only a few demonstrations of this task and from that, learn the interaction primitives (such as, moving, grasping, lifting and releasing) as well as the interaction network (collaborative assembly) for the task. To do this, we propose to enhance traditional imitation learning by using speech markers.

For imitation learning, we use Simtrack [12] which allows us to track the parts of the toolbox based on 3D models that were provided to the system. Skeleton tracking is done by using OpenPose [13] [14] [15], providing a reliable 2D skeleton estimate of the user. As robotic platform, we use the UR5 robot equipped with a Robotiq adaptive three finger gripper.

The process of retrieving the speech markers for temporal labeling is divided in two steps. First, free speech

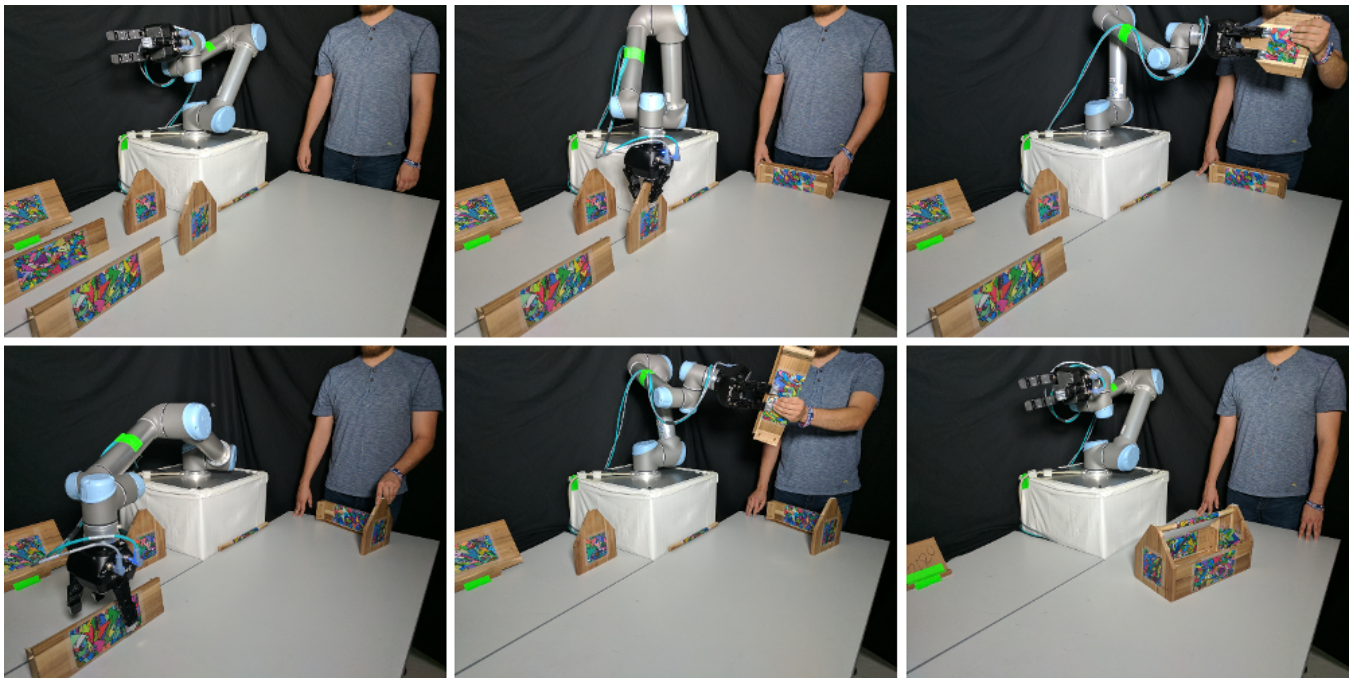

Fig. 2. This image shows an overview over the performed experiment. The task consists of assembling a wooden toolbox where five out of six parts are out of reach for the human. From left to right the image sequence shows the successful assembly of the box.

Goals: Categories and Synonyms					
human	*long side one*	*long side two*	*short side one*	*short side two*	*bottom plate*
me	long side one	long side two	short side one	short side two	bottom plate
idle	long piece one	long piece two	short piece one	short piece two	ground plate
	long edge one	lone edge two	short edge one	short edge two	bottom part
	long edge 1	long edge 2	short edge 1	short edge 2	base
	long object one	long object two	short object one	short object two	
	Part 1	Part 2	Part 3	Part 4	Part 5

TABLE I

THIS TABLE SHOWS AN OVERVIEW OVER THE PREDEFINED GOALS AND THEIR SYNONYMS. SYNONYMS WHERE INTRODUCED BY THE USERS WHERE AS THE ITALICIZED WORDS ARE USED INTERNALLY TO IDENTIFY THE GOALS AFTER THEY WHERE IDENTIFIED BY THEIR SYNONYMS.

is translated into text by using the Google speech to text API offered through Android. Second, the resulting text is passed to a Google based NLP API [16] that is trainable for various purposes. We trained it to extract the desired action (Table II) and goals (or objects of interest) (Table I). These actions and goals, together with their moment of occurrence serve as speech markers. Thus, a marker m can be expressed as a tuple $(action, goal, time)$. To train the NLP API we collected free speech samples from three users who were asked to talk to the robot while a demonstration was given. These samples were then used as examples for the NLP agent and manually labeled for for the $(action, goal)$ pairs. Also, this process outlined the synonyms as seen in Table I and II.

Markers are used to outline the completion of a subtask. Whenever a subtask is done, the collected data from the joints of the robot, the gripper and the object of interest are combined. The object of interest is outlined by the $goal$ of the tuple m. The data from all other objects are discarded since they were not important for the subtask, thus drastically reducing the dimensionality of the data from 61 to 19 dimensions. It is important to note that the human is also considered as an 'object of interest', depending on the received speech marker. The data can be represented as tuple $\mathbf{d} : (\mathbf{r}, g, \mathbf{o})$ where $\mathbf{r} \in \mathbb{R}^6$ holds the joint angles of the robot, $g \in \mathbb{R}$ is the control value of the gripper describing how far it is closed and $\mathbf{o} \in \mathbb{R}^{12}$ holds the values of the object. The dimensionality of $\mathbf{o}$ is variable, since a simple object (highlighted in green in Figure 1) is described by its 3DOF position and 3DOF rotation where as the human skeleton is described by the 2DOF positions of the wrist, elbow and shoulder position of each arm. To compensate for the different lengths, the object information is padded with zeros to also be of length 12.

Based on these data, Interaction Primitives are used to train the actions of the individual subtasks. The training is done on all 19 dimensions of $\mathbf{d}$. At run time, the necessary movements of the robot $\mathbf{r}$ and gripper g are unknown and will be generated by the IPs based on the current observation of the object of interest $\mathbf{o}$. This procedure results in one specific IP for every action and goal combination. In theory,

Actions: Categories and Synonyms			
lift	*grasp*	*move*	*release*
lift	grasp	move	release
hand over	pick	goto	free
handover	pick up	get	
give	pickup	collect	
lift up		move to	
bring		go to	
hand		moving	
		moved	

TABLE II

THIS TABLE SHOWS AN OVERVIEW OVER THE PREDEFINED ACTIONS AND THEIR SYNONYMS. SYNONYMS WHERE INTRODUCED BY THE USERS WHERE AS THE ITALICIZED WORDS ARE USED INTERNALLY TO IDENTIFY THE ACTIONS AFTER THEY WHERE IDENTIFIED BY THEIR SYNONYMS.

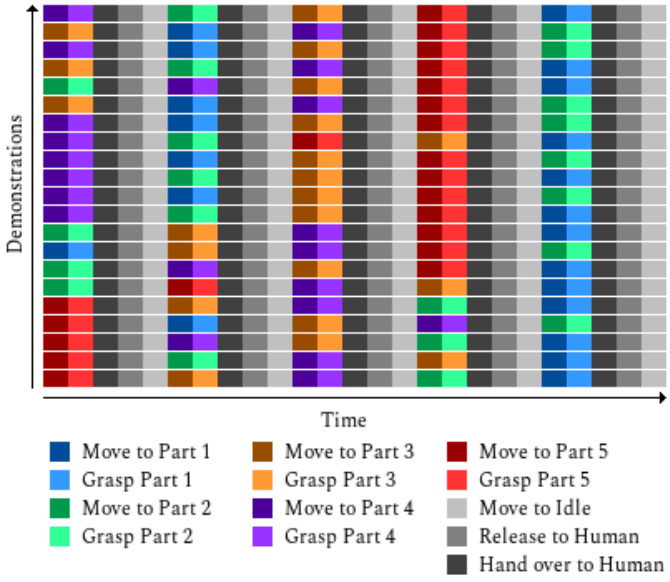

Fig. 3. Overview over the order in which the states appeared. Every square represents a certain combination of action and goal.

every combination of action and goal has exactly one IP. Depending on the training, some may be left empty (e.g. m where $action$ = 'grasp' and $object$ = 'human'). Subtasks that occurred multiple times during one demonstration receive multiple training examples from one full assembly.

When recording multiple demonstrations of a certain task, the order in which the speech markers m appear can be used to infer and abstract the task. The next section will have a closer look on how to utilize the state order to create an abstraction of the performed task, which is then allowing the robot to decide on the object of interest at run time.

Figure 4 shows the final interaction network that is based on 21 demonstrations shown in Figure 3. Due to a different order in which objects are requested by the user, each demonstration in Figure 3 can have a different order in which the task was completed. Partial ordering can be inferred for the different parts of the toolbox, separated by the handover subtask. In figure 3, colored blocks are subject to eventual partial ordering based on the given demonstrations. At run time, the interaction network is transitioned independently by the robot. When multiple choices are available, the robot decides randomly which instance of action and goal m' it takes. However, the random decision is limited by two constraints which ensure that the task can be finished. First, it does not repeat actions that are already finished and second, it makes sure that there is a path to the end. When looking at the right-most level of the interaction network, the last level involves 'part 1' and 'part 2'. The second condition ensures that one of these parts is still in an unfinished state when reaching the last level. This prevents the robot from getting stuck during the interactions. To have more or all choices in this last level, a demonstration that showed another part on the last level would have been necessary. In this scenario, the human demonstrator did not give this possibility to the robot.

III. EXPERIMENTAL EVALUATION

We performed three experiments to show the benefit of using speech markers during training in comparison to traditional methods. Furthermore we show the ability of the interaction primitives to adapt and generalize towards different environmental conditions, e.g. object locations and orientations.

The first experiment uses an environment as shown in figure 1, where the objects where placed in the same position within a margin of $\pm$4cm after each demonstration. In total, five demonstrations of the whole assembly task were provided. Based on this training, the task was executed ten times with a success rate of 80%. Failures were due to failed grasp attempts or collisions of the grasped object with other objects in the environment. This error can be explained by the different configurations of the five objects among different demonstrations. Depending on the order in which objects are collected, the pickup trajectory can be different due to more or less space in the storage area. At run time, the IPs generated an average over all demonstrated pickup movements for a particular object without considering other objects, occasionally resulting in an early side movement that leads to collisions.

The second experiment evaluates the performance of the system in the same setup as in the first experiment, but without the use of speech to refer to an object. This means that all 61 dimensions are considered by the interaction primitives since no object of interest was outlined. Clustering is done manually by pressing a button after each subtask is finished. As in the first experiment, the system received five demonstrations of the assembly task. However, ten executions without our system resulted in a success rate of 0%. This is due to the high dimensionality of the input data because of which the interaction primitives are not able to condition on the important properties for the individual task.

The third experiment evaluates the ability to generalize towards different object positions. For simplicity the experiment was only conducted with one object which was

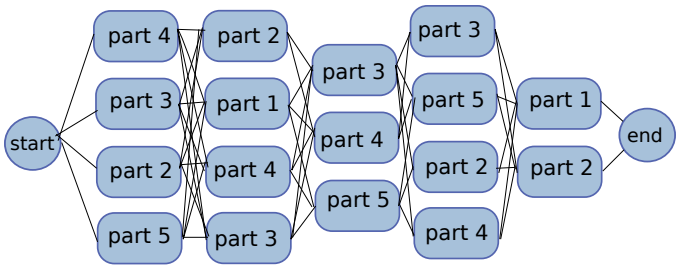

Fig. 4. This figure shows a interaction network for the assembly task including partial ordering. In every step, one box can be chosen.

randomly placed on the table as shown in Figure 5. Eight demonstrations were provided to the system outlining the outer boundaries of the table as well as some positions in the middle. This model was then evaluated by attempting to grasp the object from random positions within the trained limits, which resulted in a successful grasp in 80% of the trials. The two failed trials were due to the gripper not being able to grasp the object. In general it was noticeable that the accuracy of the grasp was worse than in the first experiments. With regards to the center of the object, the robot grasped the object with a spread of $\pm$7cm from the center where as the grasp in the first experiment only spread for about $\pm$2cm.

IV. Conclusion and Future Work

We developed a methodology to use speech during training with long interactions and our experiments showed that this can drastically improve the quality of the training since the user can outline the object of interest. This allows the robot to focus and train on the important aspects rather than getting lost in observing irrelevant details. This theory is supported by the first two experiments in which our system improved the success rate of the assembly task. A similar conclusion was also drawn in [11], but our system is able to leverage speech to ease the training and outline one of multiple objects as object of interest. Future work will focus on using broader knowledge from speech and text to enable direct feedback and contextual questions from the robot to allow more natural interactions with the system. Additionally, the removal of all objects except from one might be too harsh. The first experiment suggests that a weighted influence from all objects might be beneficial.

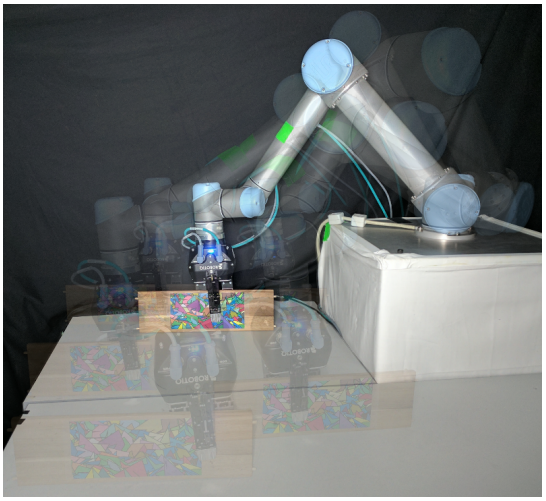

Fig. 5. Overview of the generalization capability of the interaction primitives

References

[1] H. B. Amor, G. Neumann, S. Kamthe, O. Kroemer, and J. Peters, "Interaction primitives for human-robot cooperation tasks," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2831–2837, May 2014.

[2] C. G. Atkeson and S. Schaal, "Robot learning from demonstration," in *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, (San Francisco, CA, USA), pp. 12–20, Morgan Kaufmann Publishers Inc., 1997.

[3] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

[4] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, "Robot programming by demonstration," in *Springer handbook of robotics*, pp. 1371–1394, Springer Berlin Heidelberg, 2008.

[5] C. Mericli, S. D. Klee, J. Paparian, and M. Veloso, "An interactive approach for situated task specification through verbal instructions," in *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, (Richland, SC), pp. 1069–1076, International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[6] L. She, Y. Cheng, J. Y. Chai, Y. Jia, S. Yang, and N. Xi, "Teaching robots new actions through natural language instructions," in *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, pp. 868–873, IEEE, 2014.

[7] L. She, S. Yang, Y. Cheng, Y. Jia, J. Y. Chai, and N. Xi, "Back to the blocks world: Learning new actions through situated human-robot dialogue," in *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, vol. 89, 2014.

[8] H. Cuayáhuitl, "Robot learning from verbal interaction: a brief survey," *Proceedings of the New Frontiers in Human-Robot Interaction*, 2015.

[9] M. N. Nicolescu and M. J. Mataric, "Natural methods for robot task learning: Instructive demonstrations, generalization and practice," in *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '03, (New York, NY, USA), pp. 241–248, ACM, 2003.

[10] G. Gemignani, E. Bastianelli, and D. Nardi, "Teaching robots parametrized executable plans through spoken interaction," in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, (Richland, SC), pp. 851–859, International Foundation for Autonomous Agents and Multiagent Systems, 2015.

[11] B. Akgun and A. Thomaz, "Simultaneously learning actions and goals from demonstration," *Autonomous Robots*, vol. 40, pp. 211–227, jul 2015.

[12] K. Pauwels and D. Kragic, "Simtrack: A simulation-based framework for scalable real-time object pose detection and tracking," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1300–1307, Sept 2015.

[13] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[14] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *CVPR*, 2017.

[15] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.

[16] "Api.ai."